It's not or isn't it?

Using large corpora to determine the influences on contraction strategies

Language Variation and Change, **14** (2002), 79–118. Printed in the U.S.A.

DOI; 10.1017.S0954394502141044

It's not or *isn't it?* Using large corpora to determine the influences on contraction strategies

MALCAH YAEGER-DROR, LAUREN HALL-LEW, AND SHARON DECKERT
University of Arizona

ABSTRACT

In analyzing *not*-negation variation in English it becomes clear that specific strategies are used for prosodic emphasis and reduction of *not* in different social situations, and that contraction strategies vary independently of prosodic reduction. This article focuses on the factors influencing contraction strategies that are clearly dialect related and attempts to tease out those factors that are related to register and speaker stance. First, we review background information critical to an adequate analysis of *not*-negation and *not*-contraction. We then describe the corpora chosen for the present study, the research methods employed in the analysis, and the results of the analysis. The variable under analysis is the choice between uncontracted and *not*-contracted forms and between *not*-contracted and Aux-contracted forms in well-formed declarative sentences, for verbs which permit both. We end with some suggestions for corpus composition that will enable meaningful comparisons between social situations and between speakers, or characters, within one corpus. As researchers we can assure that future corpora will permit increasingly inclusive and interesting comparative studies; we close with some suggestions for those who wish to carry out studies.

Tottie (1991) showed that there are three direct ways to express negation in English. These are shown in Table 1. She found that the vast majority of English negatives used are *not* negatives. For that reason, the present study narrows its focus to the analysis of *not*-negation. *Not*-contraction, which entered the English language around 1600 (Jespersen, 1917; Warner, 1993) or even earlier (Rissanen, 1999), has become the norm in most varieties of spoken British and American English. "British" is used as a cover term for the English spoken in England, Scotland, and Ireland and "American" as a cover term for United States and Anglophone Canadian speech. British nonfiction uses contraction less consistently

This article would not have been written without the funding of NSF#9808994 or without the LDC corpus, accessed—with much technical support and advice from Dave Graff and the moral support of Merrill Garrett—through the Cognitive Science Program of the University of Arizona. We are also grateful to Mary Finch of the Bush Presidential Library, Ron Whealan of the Kennedy Presidential Library, and the linguistics librarian at the University of Arizona, Sara Heitshu, all of whom facilitated access to the text and sound files of political corpora. Doug Biber, Crawford Feagin, Candy and Chuck Goodwin, Greg Guy, John Heritage, Chuck Meyer, Michael Montgomery, and Sali Tagliamonte provided interesting insights, as did the journal's anonymous reviewers. Any remaining shortcomings are our own.

TABLE 1. *Types of negation*

Tottie's Terminology	Examples	Sample Sentences
not-negation	*is not, isn't, 's not*	*It isn't really possible.*
no-negation	nowhere, never, nothing, nobody …	*I never did that.*
affixal negation	*imperfect, irrespective, independent, nonfunctional, disingenuous, unable …*	*I am incapable of doing it!*

Source: Tottie (1991).

than fiction (Kjellmer, 1998). Older texts use full forms more than recent texts in the same genre or register (Biber, 1988). Written registers use full forms more consistently than spoken registers (Kjellmer, 1998, and Tottie, 1991, for British English; Yaeger-Dror, 1997, for American English). Bell (1984) showed that, in declarative sentences in news reporting, contracted forms are more common in the United States than in the British Commonwealth, and it is commonly believed that the full form is more common in British than in American conversational declaratives as well.

Biber (1988), who has done the most work to compare large linguistic corpora from different social situations (or speech "registers"), showed that, if a multivariate analysis is carried out on information concerning variation in many linguistic factors, five register continua (or dimensions) can be isolated for English. Dimension 1, the statistically most significant of these, is a continuous parameter fluctuating from a more informative pole, which he referred to as "Informational," to a socially more interactive pole, which he referred to as "Involved." There is now a fair amount of evidence to support the claim that register influences contraction strategies.

Cognitive theories would project that, when given the choice, forms with full *not* retained would be favored in informative settings, since *not* carries important semantic information; we have referred to this as the Cognitive Prominence Principle (Yaeger-Dror, 1996, 1997, 2002b). It is not coincidental that contraction is the most extreme form of lexical reduction available to the English speaker, and that speakers avoid contraction in informative situations where the significance of *not* is most important.

The pattern for interactive data is also influenced by a conflicting Social Agreement Principle, which has been identified in the work of Goffman (1971), Sacks (1992), and their students; these researchers showed that repair (or remedial turns) is dispreferred in conversation, whereas supportive turns are preferred. Yaeger-Dror (1985, 1997, 2002a, 2002b) found considerable evidence that, in both spoken and written use of American English, where the speaker's purpose is informative and socially neutral, full forms of *not* are favored; *not*-contracted forms are favored when the speaker's purpose is supportively interactive and the negation is used to express a repair. Consequently, Yaeger-Dror (1996, 1997, 2001, 2002b, 2002c) found that negation has characteristics that particularly militate against a simple analysis based on one corpus with a limited register, and

that the ideal would be to access data that would enable researchers to compare the negatives used by speakers in several social situations.

Holding dialect and chronological era steady, the informative registers of British and American English (e.g., news, tutorials, or written descriptive texts) use more full forms than the interactive registers (e.g., conversations or written dialogue) (Biber, 1988; Yaeger-Dror, 1997).[1] In her analysis of *not*-negation in two corpora of newspaper prose, Westergren-Axelsson (1998) provided confirmatory evidence for variation in negative syntactic strategies, which coincides with the informative–interactive continuum. She isolated three subgenres: reporting, editorials, and reviews.[2] She also made a distinction (already isolated in Yaeger-Dror, 1996, 1997) between material inside and outside quotation marks—with more *not*-contraction in dialogue and actual conversations than in informative/narrative segments of text. Both Yaeger-Dror (1997; Yaeger-Dror, Hall-Lew, & Deckert, in press) and Westergren-Axelsson found that there is a large gap between narrative prose and written dialogue, which are presumably more informative and more (pseudo)interactive, respectively.

Both contraction and prosodic strategies in dialogue differed significantly from those in read descriptive prose and were more similar to actual interaction. Thus, when interactive rules are more relevant—whether signaled by quotation marks in print or triggered by the interactiveness of a social situation—the likelihood of *not*-contraction increases; conversely, when conveying information is primary, *not*-contraction is curtailed.

Biber and Finegan developed the ARCHER corpus specifically to permit analysis of change in time (Biber, Conrad, & Reppen, 1998). When ARCHER written corpora from different eras were compared, Dimension 1 was shown to have varied over the last few centuries; some social situations (or registers) have become more interactive, while others have become more informative. Thus, for example, while most registers have become more interactive, medical writings have become more informative (Biber, Finegan, & Atkinson, 1993:9). Both American and British journals (diaries) have become more interactive over the last 240 years; moreover, the American diaries that were initially more informative have become more interactive (1993:10).

Not only is intention of the speaker relevant (to convey information? to repair another's turn? etc.), but speaker stance (Goffman, 1981) is critical as well. In certain registers disagreement is preferred. For example, considerable evidence has now been presented to show that children express disagreement prominently in certain registers (Corsaro & Rizzo, 1990; Goodwin, 1983; Goodwin, Goodwin, & Yaeger-Dror, 2002; Hoyle & Adger, 1998; Kyratzis & Guo, 2001; Sheldon, 1996, 1998). Even among adults, an adversarial stance, which requires negation to be emphasized, is not as uncommon as early conversation literature would have us conclude (Clayman, 2002, in press; Heritage, 2002; Hutchby, 1996, 1999).

For example, comparing evidence from political debates with data from other registers, Yaeger-Dror found that interactive registers vary along this second continuum, from more supportive turns, as in conversational interactions analyzed by Sacks (1992) and Schegloff, Jefferson, and Sacks (1977), to more adversarial turns, as in political interviews, legal interactions, and debates (Yaeger-Dror,

1996, 1997, 2002a). She also found that full *not* would be retained in an adversarial stance, as in debates, but not when used by the program moderator (Yaeger-Dror & Hall-Lew, 2000). Other situations in which adults were expected to express disagreement quite emphatically (but the moderators were not) included political TV programs (Blum-Kulka, Blondheim, & Hacohen, 2002; Scott, 1998), politicians' news conferences (Clayman, in press; Heritage, 2002; Perez de Ayala, 2001), call-in programs (Hutchby, 1996, 1999, 2001), candid camera programs (Al-Khatib, 1997), and talk shows (Ilie, 1999). In such situations, unreduced *not* tokens were preferred for the participant stance.

In contrast, overt disagreement is even more tabooed for moderators or negotiators than it is for polite conversationalists (Clayman, 2002; Jacobs, 2002). It is also true that there is a growing body of evidence to demonstrate that a specific register may require an adversarial stance in one culture but not in another (Yaeger-Dror, 2002a, 2002b).[3]

Contraction is also correlated with sentence type. *Not* tokens in imperatives, as in (1), and interrogatives, as in (2), are almost categorically contracted in American English (Yaeger-Dror, 1996, 1997), even in informative written contexts. Contraction is not inevitable in these sentence types in British data, as shown in examples (2c) through (2g). In his analysis of written British corpora from the 1960s, Kjellmer (1998) found that questions were only 90% *not*-contracted. In their study of interviews with older speakers of northern and rural British dialects, Tagliamonte and Smith (in press)[4] found that only 65% of questions were *not*-contracted, although tags were categorically contracted as they were in our American sample. Moreover, Tagliamonte and Smith found that for Scots speakers the bulk of the contracted forms were used in rhetorical questions, which are more like tags, and so the percentages for full forms in questions requiring an answer were even higher and less like the American interrogatives. The interaction of dialect with sentence type must be considered as a separate issue.

(1) a. Please don't eat the daisies! (American)
  b. Don't mess with social security! (PD, St. Louis Debate, Bush 978)[5]
  c. . . . But don't just sit here slow dancing for 4 years! (PD, Richmond Debate, Perot 310)
  d. Don't go away yet! (PD, St. Louis Debate, Jim Lehrer 1130)

(2) a. Isn't she sweet?/Ain't she sweet? (American)
  b. "Oh, aren't you well? Sha'n't I bring your dinner?" (Wharton, 1911/1969, 253)
  c. Well, why should they not use the words of the original? (BNC, S1.1, 964)
  d. but . is– is that not a private library? (BNC, S3.3, 194)
  e. Is there not somewhere you can copy it up? (COLT, F, 47yrs.)
  f. Is that not scary as crap? (COLT, F, 17yrs.)
  g. Is she not wearing tights today? It doesn't look very nice. (COLT, F, 14yrs.);

For the present study the analysis focuses only on *not*-negatives in complete declarative sentences. Interrogatives and imperatives are not included in the analysis, nor are sentences that are radically elliptical.

In English language studies (e.g., Biber, 1988), as in the preceding overview, the full form is generally contrasted with the contracted form. The majority of

TABLE 2. *Full and contracted forms*

Full Form	*not*-Contracted	Aux-Contracted
He cannot	*He can't*	—
He has not done that	*He hasn't done it*	*He's not done that*
We will not do it	*We won't do it*	*We'll not do it*
We have not done it	*We haven't done it*	*We've not done it*
He is not here	*He isn't here*	*He's not here*
We are not here	*We aren't here*	*We're not here*

verbs fall into a class which only permits one form of contraction: *not*-contraction. Consonant-initial verbs which only permit *not*-contraction are referred to here as "other" verbs.

However, Table 2 shows that for some verbs there are actually two possible contracted forms. These two contraction strategies are referred to as *not*-contraction (*isn't*) and Aux-contraction (*'s not*). Note that Aux-contraction permits *not* to be unreduced; this has important implications for our hypothesis, which posits that uncontracted negatives would be more likely to occur in informative registers (Yaeger-Dror, 1997, 2002a).

The extent of variability in contraction strategies and the fact that they are subject to internal-linguistic constraints (such as preceding phonological unit or whether the sentence is declarative or interrogative) as well as dialect and register constraints provide a very interesting set of problems for analysis.

Initially it was assumed that Aux-contractable verbs were classed together because they were vowel-initial (that is, for a fairly surface structure reason). However, Lightfoot (1999:186–195) presented historical evidence that {*be*, *will*, *have*} have never functioned like other English verbs. While surface factors (e.g., whether the preceding word ends in a vowel or a consonant) are certainly relevant to choice of Aux-contracted or *not*-contracted form (Hiller, 1987; Kjellmer, 1998; McElhinny, 1993), they are not discussed in detail here.

In British dialects the relative frequency of Aux-contraction is said to increase the further north one goes (Trudgill, 1978). However, Tagliamonte and Smith (in press) found that, while the range of Aux-contraction varies more widely than it does for our American speakers, the geographical picture is much more complex than Trudgill's comments would suggest.

A number of linguistic constraints restrict variation. In Britain {*will*, *is*, *are*} are said to be contracted more often than other auxiliaries in declarative sentences (Tagliamonte & Smith, in press).[6] For American English we try to show that {*have*, *is*, *are*} are the auxiliaries most often contracted, while *will* is very rarely contracted in our corpus.

Jespersen (1917) and Denison (1999) discussed the fact that *amn't** → *an't* → *ain't*. By the early 20th century, in England *ain't* was still preferred to *aren't* as a contraction for *am not*, but was condemned as a contraction for other subjects. That is, *I ain't* was acceptable in British English, while *he ain't* was a stigmatized vernacular form (Trudgill, 1990:94). In the Irish and British vernaculars today, *an't*

or *amn't* is used in *r*-ful dialect areas, and *aren't* is used in *r*-less dialects (Bresnan, 2000; Hudson, 2000; Tagliamonte & Smith, in press; Trudgill, 1990).

Two scholars have analyzed variation in *not*-contraction and Aux-contraction among southern vernacular speakers in the United States, based on analysis of their own Labov-style interviews. Both Feagin (1979) and Hazen (1996) analyzed the ratio of different *be*-contraction forms, {*isn't, 's not, ain't*} and {*aren't, 're not, ain't*}, among southern vernacular speakers and found that there is a high percentage of {*is not, are not*} realized as *ain't* tokens for both rural and urban working-class speakers, while there is a much lower percentage of *ain't* among urban middle-class speakers. They both found that the use of *is* and *are* as verb or copula did not appear to influence contraction preferences. Thus, in southern non-middle-class vernacular, as in British vernaculars, *ain't* is used not just for *am not*, but also for *isn't*, *aren't*, and even *haven't* and *hasn't* (Feagin, 1979; Hazen, 1996).

Given that *ain't* occurs only rarely in our sample, except in some of the 19th-century literature, these findings are relevant primarily because of a conjecture made by Feagin. Looking at evidence from change in apparent time, Feagin suggested that the fact that Southern Standard speakers appear to favor Aux-contraction over *not*-contraction may stem from their wish to avoid *ain't*. She thus assumed that southern middle-class speakers use Aux-contraction for {*is not*/*are not*} more consistently than middle-class speakers from other regions. In fact, one of our goals in this study is to determine if speakers from different regions in the United States vary in the extent to which they favor (or disfavor) Aux-contraction.

Both *is not* (3) and *are not* (4) tokens can be full, Aux-contracted, or *not*-contracted.

(3) a. But that is not the only way . . . (Wharton, 1917/1998, 639; descriptive passage)[7]
b. "That's all, is it? It's not much!" (Wharton, 1917/1998, 402)
c. This is not mud slinging. This is fact! (PD, Richmond Debate, Perot 224)
d. Now, it's not the Republicans' fault, . . . (PD, Richmond Debate, Perot 82)
e. But it isn't going to get the job done, . . . (PD, Richmond Debate, Perot 594)
f. Runway 3-3 is not available this morning. (ATC, Boston, 20295)
g. If it's not, I'll uh– I'll have to check. (ATC, Dallas, 31457)
h. . . . but the frequency isn't -uh– that good. (ATC, Boston, 42477)

(4) a. All the nuclear weapons are not dismantled! (PD, Richmond Debate, Perot 224)
b. You know, we're not under oath at this point! (PD, Richmond Debate, Perot 164)
c. Everybody cares if people aren't doing well! (PD, Richmond Debate, Clinton 448)
d. Replies are not received from several flights. (ATC, Boston, 38524)
e. Alright, they're not working then. (ATC, Dallas, 14890)
f. . . . We don't know because inspectors aren't in. (PD, Bush/Gore, Debate 3)

As already stated, the present study focuses on the choice between Aux-contracted and *not*-contracted declarative tokens of {*is not, are not*} which do permit variation in American English. *Ain't* occurs too rarely in the corpus analyzed here to be discussed further.

We start with the hypothesis that prominent *not* is preferred in informative situations and dispreferred in interaction (Yaeger-Dror, 1996). Meaningful conclusions about dialect influence on contraction strategies can only be drawn when there is ready access to a large corpora of transcribed speech coded for a range of sociolinguistic variables (age, sex, region, social class) as well as register. The present study makes use of a corpus which permits a pilot study of such dialect variation and compares the results with those from data in other registers.

Today, there are several large transcribed corpora which permit analysis of lexical and morphological variation, and studies based on those corpora have been published (Biber, Conrad, & Reppen, 1998; Biber, Johansson, Leech, Conrad, & Finegan, 2000; Johansson & Oksefjell, 1998; Kennedy, 1998). The present study analyzes data from several large corpora which are now available. The primary goal is to analyze variation in contraction strategies, and a secondary goal is to determine how feasible it is to compare data from corpora which differ in multiple ways simultaneously.

Earlier studies of the use of negatives have shown that sentence type, dialect, time, social situation, and speaker stance all influence contraction strategies. Certainly the full form is more common in interrogatives and imperatives in British English than in American English, where contraction has been found to be almost categorical even in writing (Yaeger-Dror, Hall-Lew, & Deckert, ms.). The consensus is that in declarative sentences contraction has become more acceptable in 20th-century written texts. However, we show that even today scripted texts (like read news) use full form more consistently than unscripted informative texts (like Air Traffic Control or academic data). We project that Aux-contraction for {*is not*/*are not*} is higher in those registers which permit more full form (i.e., informative and adversarial stance registers), and that dialect is a relevant parameter in the United States as well as in the United Kingdom.

One purpose of the present study is to determine whether, holding register and sentence type steady, regional dialect is a factor for not-contraction. Another purpose is to discover whether it is possible to analyze variation in register, stance, and dialect simultaneously to determine which has the strongest influence and to see whether the corpora presently available permit any viable multivariate analysis.

THE PRESENT CORPUS

Various corpora are analyzed to attempt a meaningful comparison of {*is not, are not*} contraction strategies in declaratives for different registers and among speakers from different regions. Our goal here is to demonstrate the influence of dialect and register on the choice between Aux-contracted and *not*-contracted forms. In the course of the preliminary discussion, we characterize the different corpora used; the contraction percentages for the "other" verbs in these corpora are noted. The problems facing researchers who wish to do a comparative survey using ready-made corpora should become obvious.

Table 3 provides a list of the corpora consulted for this study, when and where they were collected, and the number of words in the corpus. The Appendix at the

TABLE 3. *Corpora included in this analysis*

Text Type	Source	Date	Region	Number of Words
Informative!	ATC	1980s	ne/DC/w	—
Radio: Boston News	NPR	1980	ne:MA	54,739
Marketplace	USC	1996	Los Angeles	—
ftf: Lectures	MICASE	1995f	nc:MI~	222,000
Student Presentations	″	″	″	83,027
Seminars	″	″	″	34,982
Defenses	″	″	″	48,596
ftf: Q/A	Kennedy	1961–62	ne:MA	93,545
	Nixon	1969–74	w:CA	50,166
	Ford	1974–76	nc:MI	37,710
	Carter	1977–80	s:GA	68,664
	Reagan	1981–86	nc:IL	52,272
	Bush	1989–91	ne/CT	30,727
	Clinton	1993–99	s/w:AR	35,961
Informative News	*G/M*	1999	CDN	49,446
Informative News	*NYT*	1999	US	??
Inf. §s	*NYT*	1997–99	US	33,689
Inf. §s [1st p]	*NYT*	1997–99	CDN[a]	23,426
Book Reviews	*NYT, NYRB*	1999	US	29,688
Book Reviews	*G/M*	1999	CDN	38,877
Literary §s[1st p]	*NYT*	1997–	CDN[a]	18,475
Literary §s	*NYT*	1997–	CDN[a]	10,784
Literary §s	*NYT*	2000	US	28,000
Literary Texts	Jane Austen	1814	UK:S	159,911
	C. Bronte: *J.Eyre*	1846	UK:N	186,000
	E. Bronte: *Wuther*	1847	UK:N	116,700
	A. Bronte: *AG*	1848	UK:N	169,000
	Hawthorne	1850	US:MA	83,688
	Stowe	1852	US:ne	182,450
	Gaskell: *N&S*	1853–54	UK:N	284,000~
	Eliot: *Mill*	1860	UK:MID	45,000
	Collins: *Woman*	1860	UK:London	252,160
	Dickens	1861	UK:London	187,400
	Trollope: *PhF*	1869	UK:S	126,750
	Alcott: *LittleMen*	1871	US:MA	185,800
	Trollope: *EuD*	1873	UK:S	272,500
	Twain: *Huck/Tom*	1876/1884	US:MO/LA	240,000~
	Twain: *Abroad*	—	US:MO/LA	—
	Henry James	1880	US:NY/London	64,000
	Twain: *Yankee*	1889	US:MO/CT	120,700
	Hardy: *Tess*	1891	UK:S	150,000
	Chopin	1899	US:St. Louis	90,700
	Glasgow	1904	US:VA	135,300
	Cabell: *Hour*	1909	US:VA	56,000
	Wharton: *EFrome*	1911	US:NY/MA	31,274
	Wharton: *Summer*	1914	US:NY/IL	57,437
	Maugham	1915	UK:Kent	76,000
	Cather: *Lark*	1915	US:NE	152,700
	Woolf: *N&D*	1919	UK:London	167,300

continued

TABLE 3. *(continued)*

Text Type	Source	Date	Region	Number of Words
	Cather: *Prof*	1925	US:NE/NY	62,000
	Cleary	1950	US:OR	22,597
	Tyler	1988	US:NC/MD	65,932
	Beattie	1990s	US:DC/ME	5,100
	Keillor	1985	US:MN	13,000
Interactive, phone	SWB by region	1980s	US	339,000
			ne/n	56,000/
			n mid	56,000
			nyc	56,000
			s	56,000
			s mid	56,000
			w	56,000
Interactive, phone	Call Home	1980s	US	144,000
Interactive, ftf	Upholstery Shop	1971	NYC	32,500
Interactive, ftf	Segrin, Supportive	1996–97	midwest	74,400
Interactive, ftf	Segrin, Remedial	1996–97	midwest	140,400
Interactive, ftf	COLT: adult	1993	UK[teen]	500,000?
	teenager	″	″	″
Adversarial, phone	T/L	1997–	n/w	16,962
Adversarial!, ftf	PD: Kennedy x4	1960	ne:MA	18,032
	PD: Nixon x4	1960	n/w:CA	18,648
	PD: Ford x3	1976	nc:MI	12,854
	PD: Carter x4	1976	s:GA	15,242
	PD: Reagan: cmb.	1980	nc:IL	25,548
	M/L: Mecham	1988	w:UT	1,260
	M/L: Babbitt	1988	w:AZ	475
	PD: Bush x3	1992	ne:CT	15,170
	PD: Perot x3	1992	s/w:TX	13,296
	PD: Clinton x3	1992	s/w:AR	14,450
	PD: GWBush x2	2000	w:TX	13,015
	PD: Gore x2	2000	s:TN	13,143
	MD: Tucson	1999	w:AZ	7,670
	MD: McCasson	1999	w	1,782
	Ont. Primary	1995	CDN	18,000

Note: See the Appendix for more elaborate discussion of the corpora.
[a]In the NYT corpus, Canadian authors were isolated from United States authors.

end of this article clarifies the abbreviations used to refer to the different corpora, describes the corpora a bit more fully, and lists a URL where the corpus, or at least a description of it, can be found online.

Written registers

Informative journalistic prose from the United States and Canada was collected directly from the web. The advantage of downloading one's own text corpus is that one can choose a very narrow, clearly defined register and verify the native dialect area of specific journalists or authors included in the sample. The Oxford Text Ar-

chives, the British National Corpus (Aston & Burnard, 1997), and the Linguistic Data Consortium (LDC) all have journalistic text files, but they do not permit the analyst to code for dialect or register. The massive news files include wire data (e.g., from AP or Reuters) which preclude the tracking of dialect information. They also merge files from various sections of newspapers which are dissimilar in register and stance. Downloading one article at a time is much more time-consuming, but part of that time is spent verifying where the journalist is from[8] and determining what register is being used.

Two sources for American journalistic prose were analyzed: scientific articles from the *New York Times* (NYT) or the *New York Review of Books* (NYRB) for northeastern United States speech and from clearly Toronto born and raised journalists at the *Toronto Star*, *Macleans*, or *Globe and Mail* (henceforth collectively designated as G/M) for Canadian speech.[9] These were supplemented with first chapters (or §1) of nonfiction books from the NYT for northeastern United States authors and from the NYT and G/M for Ontario authors.

Book reviews were also collected from the same sources, with the same attention paid to where the reviewer was said to have been raised. For the most part, informative-style reviews of informative texts (biographies, science books) were chosen from the NYT and NYRB for northeastern United States authors and from the NYT and G/M for Ontario authors. Downloading one review at a time, it was also possible to verify the journalist's region of origin and his/her stance vis-à-vis the topic.[10] The comparison of NYT and G/M corpora revealed that both region and register had an impact on contraction.

Prose was also collected from the web. Older classical literary prose could be contrasted with more recent literary prose, which was only available by scanning the data[11] or by taking the first chapter samples available from the NYT and G/M. As with the journalists, authors' biographies were available on the site itself or were found with google.com. Narrative was isolated from dialogue.

British authors' use of contraction was contrasted with American authors' usage. Dialect area was determined (as closely as possible) not just for the authors but for their characters as well. It is obvious in certain cases that the authors used very different contraction strategies for some characters than for others, and that, just as descriptive prose and dialogue must be isolated, ultimately each character's dialogue should be coded separately for variables like contraction, for which generation, class, and dialect area might be critical factors in the analysis.

Three late 20th-century texts (Cleary, 1968; Keillor, 1985; Tyler, 1988) were scanned to provide data that could not otherwise be accessed. The fact that children are known to use aggravated disagreement more consistently than adults entails that children's literary dialogue should be studied. While Cleary (Ramona) or Rowling (Harry Potter) may soon be available on the web, scanning the Cleary text was the only way of getting child dialogue into the corpus.[12] The other two authors were chosen to provide 20th-century adult dialogue from specific areas of the country. In addition, all three authors' characters come from roughly the same region as the author, so contraction strategies would not be influenced by the authors' assumptions about speakers from other areas.

Semi-scripted spoken informative prose

National Public Radio Broadcast News (NPR) is an LDC corpus which includes over seven hours of news stories read by seven (4 males, 3 females) FM radio news announcers at a Boston radio station. This corpus was chosen for being representative of informative spoken registers (Yaeger-Dror, 1997); it has been used in other studies of informative speech as well (Hirschberg, 1993). We assumed that, just as the announcers' phonology would be relatively NPR standard (Dumas, 2001; Yaeger-Dror, 1991), the contraction strategies would reflect a relatively standard New England strategy for contracting in written informative prose, although there was no information available on the scriptwriter's background (Ostendorf, personal communication, Jan. 2001). The orthographic transcripts were generated by hand and included indications of where a speaker took a breath. A representative segment of this corpus was concordanced and analyzed for the present study.

Marketplace, as its name suggests, is an economics news program produced by USC Radio in Los Angeles, a division of the University of Southern California. USC Marketplace Speech and Transcripts, recorded in 1996 on site at the University of Southern California, contains 50 hours of broadcasted economics news programs and their transcripts. A CD-ROM release of the USC Marketplace Broadcast News Corpus is published by the LDC. The moderator-journalist is from Maine. A representative sample was analyzed.

Air Traffic Control Data (ATC) is an LDC corpus which includes 70 hours of voice communication traffic between controllers and pilots. The audio files are of continuously monitored data of a single FAA frequency for one to two hours. Full transcripts are provided for each audio file. Data were collected at Dallas Fort Worth (DFW), Logan International (BOS), and Washington National (DCA) by Texas Instruments; a CD-ROM was produced by the National Institute of Standards and Technology and distributed by LDC. For the present study a representative subset of those conversations was analyzed. While speech from the three airports does not appear to be identical, the variation does not appear to be regional, but idiosyncratic.[13]

Michigan Corpus of Academic Spoken English (MICASE) is a product of the English Language Institute (ELI) at the University of Michigan, started in 1997. The online concordance uses a customized search engine to consider approximately 23 transcripts (totaling 222,100 words) and is constantly being supplemented. Information on social stratification and other relevant variables (e.g., gender, age, broad disciplinary area, and speech "event type") is included with each token. Some event types for which data are already available on the web include large and small lectures, seminars, student presentations, and dissertation defenses. Note that this permits analysis of variation in the informative–interactive continuum. Lectures are more informative and discussions or defense speech events are more interactive. As more of the MICASE coded data come online the corpus will fill a niche for speech situations which are not otherwise available. Unfortunately, while age and professional standing are coded factors, dialect area is not among the speaker attributes coded, since many of the speakers coded their dialect background merely

as United States, and there is no funding available to return to collect more information even from the professors who are still at the University of Michigan (Swales, personal communication, March 2001). In addition, the online concordance cannot be manipulated to permit the reader to see a larger context; thus, speaker stance must be inferred from the evidence in a single turn at talk.

News Broadcast Question/Answer Sessions (Q/A) were included to permit the analysis of a register where negatives are used informatively. News conferences for which transcripts were available were gathered from various presidential archives, along with the sound files of those conferences. These were supplemented with the political debates of the presidents for whom data were available.

These files (from Kennedy to Clinton) permit analysis of linguistic and register changes over the last 40 years. One advantage of using presidential recordings is that the speakers' backgrounds are a matter of public record, and various registers for the same speakers are available.[14] Although the only audio data from the presidential archives analyzed to date consist of the political debates and the Q/A sessions from news conferences, CDs and transcripts of political orations, fireside chats, town meetings, and interviews are also available for many of the presidents, as well as sound files and transcripts of face-to-face and telephone conversations of various sorts (e.g., millercenter.virginia.edu/recordings.html). On the other hand, while even the LDC tapes must be checked for accuracy (Picone, personal communication, Nov. 2000), presidential tapes are particularly prone to well-intentioned tampering by those who wish to present our highest executive as conforming to the transcriber's standard of proper formal English (Whealan, personal communication, March 2001),[15] not to mention the mishearings inevitable in any transcription endeavor (Stern, 2000a, 2000b). For the present analysis, informative replies made by the president during news conferences were compared with adversarial replies made during a political debate.

Unscripted supportive conversational interactions

Switchboard (SWB) is an LDC corpus of 2,400 telephone conversations among 543 middle-class speakers (302 male, 241 female), gathered by Texas Instruments from all areas of the United States; each conversation lasted between 5 and 10 minutes. A computer-driven robot operator system handled the calls, selecting and dialing presumably unknown callees to take part in a conversation on a mutually agreed upon topic and recording the speech. No two speakers would converse together more than once and no one spoke more than once on a given topic. In the real world we do have conversations with strangers on topics that are only tangentially related to our lives, where the topic and information transfer are relatively important, as is sociability, which in our society entails a preference for agreement (Pomerantz, 1984; Schegloff, Jefferson, & Sacks, 1977); however, sociolinguistic analysis has not previously tapped data from such a speech register.

This corpus has the advantage that the dialect region each speaker grew up in (at least until age 10) is clearly specified, as is his/her sex, education, and the area code each speaker was dialing from. The subcorpus used here was limited to 600 conversations between people who claimed to be from the same dialect area.

It is important to note here that the SWB speakers from all regions followed the Social Agreement Principle (Schegloff, Jefferson, & Sacks, 1977) so closely that the corpus could be taken as a caricature of social amity. While in most friendly conversations one or two negatives in 100 tokens are used supportively, in the SWB conversations 25% of negatives are used supportively (Yaeger-Dror & Hall-Lew, 2000). In an independent analysis, Jefferson (2002) found that American speakers (from the Santa Barbara Ladies and the Newport Beach corpora, as well as from medical interactions) use *no* supportively only in very emphatic cases, although British speakers can use it as a mere acknowledgment token (otherwise known as a continuer). Her analysis supports our conclusion that the SWB corpus of interactions represents an emphatically supportive register for hyper-polite interactions.

The Bergen Corpus of London Teenage Language (COLT) is the first large English corpus focusing on the speech of teenagers. It was collected in 1993 and consisted of the spoken language of 13- to 17-year-old teenagers from different boroughs of London. The complete corpus, half a million words, has been orthographically transcribed and word-class tagged and is a constituent of the British National Corpus. Recently, access was granted to concord 151 texts online. The search program can show the distribution of an item in relation to factors such as the speaker's age, sex, socioeconomic class, location (inside London), and so forth. Most of the interactions were among (well-acquainted) teenagers, with the rest being between teens and their adult acquaintances.

The CHILDES Corpus (CHILDES) is an aligned corpus of conversations, which is available free online. We chose a subset of family interactions between parents and children in the Pittsburgh area. Approximately one hour of these conversations was concordanced using the internal system software.

The Segrin Corpus was collected by Chris Segrin at a large midwestern university to study variation in techniques for expressing both support for and complaints about one's partner (Flora & Segrin, 2000). Couples who had volunteered to take part in a quasi-therapeutic interaction were divided into those who were already married[16] (65 couples) and those who were dating (65 couples). Each couple took part in a supportive interaction—telling their partner what the points were that they appreciated the most—and in a follow-up situation where they were directed to tell their partner what "I wish you were more . . .": that is, what they felt caused the problems in their relationship. The data were recorded, but there was no outsider in the room with the speakers. The directions were on an audiorecording and on written note cards, and each response was to take three and a half minutes. The majority of the speakers were Caucasian (88%) college students (60%), with approximately 80% of them from the Kansas area (Segrin, personal communication), but they were not asked to provide information on their dialect background. For the purposes of this analysis, the data were divided into a SCS (Segrin Couples Supportive) corpus and a SCR (Segrin Couples Remedial) corpus.

The Tripp/Lewinsky Conversations (T/L) were transcribed by the Starr Investigation; transcripts and sound files have been posted on the web by ABC. Approximately two hours of conversations between Linda Tripp and Monica Lewinsky from the 1990s were concorded from those transcripts. (They are from New Jersey and California, respectively.) While these conversations are ostensi-

bly supportive, the degree to which the Tripp half of the conversations can be classed as supportive may be related to her ability to dissimulate.

Unscripted adversarial interactions

Political Debates (PD) were included to permit the analysis of an adversarial register where negatives and disagreement in general are actually preferred. Debates for which transcripts were available were gathered from various presidential archives, as well as debates between (then) local politicians and the primary debate for the 1999 Tucson Mayoral Election, which was collected primarily to add female speakers to the corpus, since three of the four Democratic aspirants were women. These were contrasted with the Q/A sessions from news conferences, as discussed previously.

While specifying the corpora to be used for this analysis, this section has also suggested the degree to which megacorpora can be relied on to facilitate analysis of data gathered in contrasting speech situations (registers) and the extent to which the analyst must be wary of downloading corpus data with many registers, dialects, and other factors varying simultaneously.

ANALYSIS

Where possible, transcripts of these corpora were put through the Concorder program (Rand, 1997); MICASE, CHILDES, COLT, and some of the Oxford Text Archives data were analyzed using corpus-specific online concordance programs. All tokens of *not*-negation in each corpus were tabulated and coded for specific potential environmental influences.

The present discussion focuses on the influence of dialect and speaker stance (interactively neutral or along the continuum from supportive to adversarial) on the choice between *not*-contracted and Aux-contracted forms for those verbs which permit both. Only data from well-formed declarative sentences are discussed.

Full or contracted forms in the corpora

First, however, we review the contraction percentages that were found for the "other" verbs. Considering the choice between full form and *not*-contracted form for the data analyzed here, Table 4 shows that the written corpora are most likely to use the full form. Table 4 also shows that Ontarian authors can be compared with United States authors from the northeast. Full form is generally 20% less common in the Ontario corpus than in the United States corpus for any given register. In the United States corpus, book reviews are 10% less likely to use full form than the corresponding informative articles or chapters from nonfiction, but this difference is not significant. There is a clear split between these scripted forms: 71% to 86% full form for journalists from northeastern United States, but 40% to 63% full form for Ontarian journalists, while literary book chapters have the lowest percentage of full forms.

TABLE 4. *Comparison of percentages of* not*-contracted data in different informative corpora*

Source	Date	Region	% Full Other Verbs		*haven't* % of All Contractions	{*isn't,aren't*} % of All Contractions	
			Narrative	Dialogue		Narrative	Dialogue
G/M journalist	1999	CDN	63	26	*	78	25*
NYT journalist	2000	US	86	21	*	*	13*
NYT §inf; 1st	1997–99	CDN[a]	43	—	100*	*	
NYT §inf; 1st	1997–99	US	52	15	100*	*	83
NYT §inf; 3rd	1997–99	US	85	30	100*	*	
G/M: bk rev	1999	CDN	53	60	100*	77	*
NYT: bk rev	1999	US	71	58	100*	*	
NYT §lit; 1st	1997–	CDN[a]	40	4	100*	*	
NYT §lit; 3rd	1997–	CDN[a]	58	9	100*	*	
NYT §lit; 3rd	1998	US	42	19	100*	55	50
ATC	1980s	ne/DC/w		36	33*		11
NPR	1980	ne:MA		49	94		69
Marketplace	1996	ME: Los Angeles		17	100*		67/63
MICASE	1995f	nc:MI~		17	100		13
″	″	″		10	100		6
″	″	″		11	100*		45
″	″	″		2	100*		30
Kennedy	1961–62	ne:MA		60	84		46
Nixon	1969–74	w:CA		76.	*		100
Ford	1974–76	nc:MI		34	100		70
Carter	1977–80	s:GA		40	17		3
Reagan	1981–86	nc:IL		18	91		43
Bush	1989–91	ne:CT		14	90		42
Clinton	1993–99	s/w:AR		24	72		22

[a]In the NYT corpus, Canadian authors were isolated from United States authors.

TABLE 5. *Comparison of percentages of* not*-contracted data in the different literary corpora*

Source	Date	Region	% Full/Other Verbs		*haven't* % of All Contractions	n{*isn't,aren't*} % of All Contractions	"{*isn't,aren't*} % of All Contractions
			Narrative (n)/Dialogue (")		Narrative/Dialogue		
British							
Austen	1814	Hampshire	100	99	100*	0*	0*
C. Bronte: *J.Eyre*	1846	York	99	81	*/*	0*	13
E. Bronte: *Wuther*	1847	York	81	41	*	0*	18*
A. Bronte: *AG*	1848	York+	100	53	*	*	43
A. Bronte: *Tenant*	1853	York+	96	54	*	*	36
Gaskell: *N&S*	1853–54	York	99	56	*[rural]	0*	8
Eliot: *Mill*	1860	Warwick	75	2	*/78*	0*	66
Collins: *Woman*	1860	London	96	48	*/100*	0*	17
Dickens	1861	London	87	30	*	0*	4
Trollope: *PhF*	1869	PubSchool	100	72	*/90	80*	50
Trollope: *EuD*	1873	Bloomsbury	96	32	*/93	*	76
Hardy: *Tess*	1891	Dorset	100	41	*/33	*	40
Maugham	1915	Kent	95	5	*	*	25
Woolf	1919	Bloomsbury	95	4	*	*	35

American							
Hawthorne	1850	MA	100		*	*	
Stowe	1852	MA	100		*/100	97	
Alcott: *LittleMen*	1871	MA	95	13	*/100	*	92
Twain: *Huck/Tom*	1876/1884	MO/LA	4	4	*	*	*
Twain: *Abroad*		MO/LA	1	1	*	*	*
Henry James	1880	NY/London	95	20	*/100*	*	71
Twain: *Yankee*	1889	MO/CT	41	52	*/*	95*	82
Chopin	1899	MO:St. Louis	99	18	*	*	93
Glasgow	1904	VA		12	97	*	47
Cabell: *Hour*	1909	VA	100	93	*/100*	*	*
Wharton: *EFrome*	1911	NY/MA	96	5	*/100*	0*	38*
Wharton: *Summer*	1914	NY/IL	96	8	*/100*	0*	0*
Cather: *Lark*	1915	NE	99	1	99	0	46
Cather: *Prof*	1925	NE/NY	27	10	*/82	*	23
Cleary	1950	OR	98	42	100*	*	73
Tyler	1988	NC/MD	16	3	100*	0	11
G. Keillor	1985	MN	18	0	*	*	*
A. Beattie	1990	DC/ME	38	21	100*	55	47

Note: For narrative descriptive passages, (n) is assumed to be informative, while dialogue (“) is initially assumed to be interactive and pseudo-supportive, except for children’s literature. *denotes fewer than 5 tokens in the cell.

TABLE 6. Not-*contracted data in three British corpora, to compare with the British literary data*

Source	Date	Region	*haven't* % of All Contractions	{*isn't,aren't* } % of All Contractions
Tagliamonte[a]	2000	Cumnock, Scotland Cullybackey, Northern Ireland	100	0–3
Tagliamonte	2000	Devon	97	57
Tagliamonte	2000	York	100	44
COLT: adult	1993	London Urban	89	8
teenager	1993	London/Cockney	100	14
BNC/leisure	1990s	North UK	92	33
BNC/leisure	1990s	Midland	91	26
BNC/leisure	1990s	South UK	96	25

[a]Refer to note 4.

The scripted spoken news from Boston in the 1980s was no less likely to use full form (49%) than the Toronto book reviews. The older news conferences were much more scripted than more recent conferences (Reeves, 2001; Whealan, personal communication). The early scripted presidents used full form as consistently as the NYT and more consistently than any of their successors. There was a clearer split between these scripted forms (60% full form for Kennedy and 76% for Nixon) and informative but unscripted patterns. Ford and Carter retained full form 34% to 40% of the time; more recent presidents retained full form even less often. The air traffic controllers, for whom full form carries the most critical information, retained full form approximately 36% of the time, and MICASE lecturers and the economics reporters interacting online retained full form only 17% of the time.

We conclude that the more carefully scripted speech is, the more likely full form is to be used. ATC data retained full form more consistently than one would project on the basis of this criterion alone. However, the information conveyed in ATC is doubtless an additional factor.

Table 5 shows that in literary narrative prose the contraction of the "other" verbs was not acceptable (except to Mark Twain) much before the late 20th century, although it was more common in dialogue by the mid-19th century. Thus the conventions for how informative the literary prose register should be is clearly changing over time. Table 6 presents the data from the British corpora.

Table 7 shows that in supportive interactive corpora full forms almost never occur, while Table 8 shows that they are most common in the early scripted debates, although it is difficult to determine the degree to which the percentages are influenced by register variation (from more to less scripted) rather than interactive intent. Intent is much more likely to have an isolable input on prosodic variation, as has been shown elsewhere (Yaeger-Dror, 2002b; Yaeger-Dror, Hall-Lew, & Deckert, in press).

TABLE 7. *Comparison of percentages of* not*-contracted data in the United States supportive interactive corpora*

Source	Date	Region	% Full Other Verbs	*haven't* % of All Contractions	{*isn't,aren't*} % of All Contractions
CHILDES	1980s	Pittsburgh	0	*	33
SWB by region	1980s	US			
		northeast	2	100*	27
		n midlands	4	96	24
		NYC	9	95	27
		south	4	96	12
		s midlands	5	96	15
		west	2	96	13
Upholstery Shop	1971	NYC:AAVE & ethnic	1	100*	29
Segrin Remedial	1996–97	KS	2	*	11
Segrin Supportive	1996–97	KS	2	*	10
Lewinsky	1997–	CA	5	100*	6
Tripp	1997–	NJ	19	100*	6

In short, where there is a choice between uncontracted and *not*-contracted realization, *not* is more likely to be uncontracted when in print or scripted readings than in speech. There is a preference for uncontracted *not* in both informative prose and in scripted informative or adversarial presentation. In conversations, or even scripted interactive data (like dialogue), the percentages of full forms is much lower. We suggest that this occurs because even a fictive preference for agreement (or Social Agreement Principle) outweighs the Cognitive Prominence Principle. The fact that Ontarian journalists use full form 10% to 20% less frequently than NYT journalists does not support our expectation that the British-origin provinces (Ontario) would use full form in print more than the crass Yanks (as Bell's results might have projected). Apparently, the Ontarians (and their editors) are more American than we think!

The question we explore next is whether, when there is a choice between Aux-contracted and *not*-contracted forms, a speaker will follow the same pattern. That is, where the situation is informative, are Aux-contracted forms more common than *not*-contracted forms? In friendly conversations, do *not*-contracted forms predominate? Or is situational variation altogether swamped by dialect variation? Similarly, when speakers present an adversarial stance, do they use more Aux-contraction?

Analysis of Aux-contracted vs. not-*contracted forms*

In theory, all verbs beginning in a vowel, semivowel, or *h* permit Aux-contraction, while the "other" verbs can only be *not*-contracting. The comparison should be made from situationally similar conversations for which the regional background of the speakers is known. Each verb+*not* combination was analyzed separately, because it became clear that all Aux-contractable verbs could not be analyzed together.

While *'s not* for *hasn't* did not occur in our data or in Gasparrini's (2001) or Tagliamonte and Smith's (in press) British data, *'ll not* was more frequent than *won't*: in Tagliamonte and Smith's Scottish subcorpus (10% *won't*) and in Wheatley Hill (30% *won't*). Similarly, Kjellmer's (1998) study of a British written corpus showed that Aux-contraction was high for *will*, but in this United States corpus and in Gasparrini's analysis of BNC urban leisure data, Aux-contracted tokens such as the one in (5) were too rare to permit analysis.

(5) *if he– um , someone invites him to lunch, he'll, he 'll not go* . . . (SWB3425A, 6955)

Aux-contracted *have not*, as in (6), is more common in our database than in the English corpora discussed in the literature.

(6) a. *If you're asking about the ILS, we've not had any problems with it*. (ATC, dfw7578)
b. . . . *have not heard of it any problems of it flooding. I've not heard of any particular problems this time*. (SWB3393A, NY, 605375)
c. *I've not gotten a chance to work with it* . . . (SWB3626B, NMid, 981625)

Nevertheless, when all tokens in a given corpus are compared in the tables, *not*-contraction of *have not* is still high. Compare the percentages of *not*-contraction out of all possible contractions that took place in declarative sentences for the corpus specified. The column for *haven't* presents the number for *haven't* divided by the number for *'ve not*+*haven't*: that is, the *not*-contractions out of all relevant contracted forms. (Again, except for a few of the literary texts, there are very few *ain't* used to mean *haven't*.) In this way, we minimize the influence of situation, which the inclusion of full forms would add. In fact, while *'ve not* occurs more than *'ll not*, Aux-contracted percentages for *have not* in the United States corpora are so low that there is no reliable way to judge what factors influence the choice of Aux-contraction over *not*-contraction. When the number of *not*-contractions out of all contracted *have not* forms was computed, the percentages varied from 95% to 100% *not*-contracted. Neither dialect region nor register appears to influence the likelihood of *'ve not*. Percentages appear to be too consistent across dialects and registers to permit any conclusions to be drawn.

A quantitative analysis of {*is not*, *are not*} is more likely to provide meaningful results because the present tense is more heavily used in interactive registers (Biber, 1988), and so there are more tokens of *is not* and *are not* in any given corpus. All variants of {*is not*, *are not*} were concorded and tabulated. Although some earlier studies differentiated between copula and full verb, neither Feagin (1979), nor Hazen (1996), nor our preliminary analysis appeared to reveal any pattern connected with this distinction, and it was omitted.[17] It is important to recall that in the tables and figures the percentages are for *not*-contracted tokens {*isn't aren't*} out of all contracted tokens for *is* and *are*.[18]

Hiller (1987), McElhinny (1993), and Tagliamonte and Smith (in press) all found that a preceding vowel statistically favors Aux-contraction of {*is not*/*are not*}, but Kjellmer's (1998) evidence from the LOB British written corpus did not. Preceding part of speech has also been proposed as an influence on *not*-contraction strategy. Our research group is more inclined toward the hypothesis that prosody and preceding part of speech influence contraction strategy,[19] and so an analysis of preceding context is incorporated into the ongoing analysis of prosodic variation. This article focuses only on the dialect and register factors that influence contraction strategies.

Informative corpora. Our hypothesis would suggest that Aux-contraction should be favored over *not*-contraction in informative tokens, where *not* should be unreduced. However, in all the written (NYT, G/M) and heavily scripted registers (NPR, Marketplace) *not*-contraction is actually quite high. Those text corpora with sufficient contraction to permit analysis are 50% to 83% *not*-contracted.

The NPR data of radio news scripts are also approximately 70% *not*-contracted, as is the main Marketplace news programming. Presidents with heavily scripted Q/A sessions (Reeves, 1993, 2001; Whealan, personal communication) are in the same range, but some of the more recent and less scripted segments of presidential news conferences permit much more Aux-contraction.

TABLE 8. *Comparison of percentages of* not-*contracted data in the different adversarial corpora*

Source	Date	Region:State	% Full Other Verbs	*haven't* % of All Contractions	{*isn't*,*aren't* } % of All Contractions
PD: Kennedy x4	1960	ne:MA	50	100*	35
PD: Nixon x4	1960	n/w:CA	57	100*	88
PD: Ford x3	1976	nc:MI	50	0*	67
PD: Carter x4	1976	s:GA	25	100*	0*
PD: Reagan : cmb.	1980	nc:IL	41	*	80
M/L: Mecham	1988	w:UT	11*	*	86
M/L: Babbitt	1988	s/w:AZ	0*	100*	0*
PD: Bush x3	1992	ne:CT	14	100*	23
PD: Perot x3	1992	s/w:TX	8	100*	6
PD: Clinton x3	1992	s/w:AR	17	100*	0
PD: GWBush x2	2000	s/w:TX	10	100	23
PD: Gore x2	2000	s:TN	25	100	13
Ont. Primary	1995	CDN:Ont	36	80*	55*
MD: Tucson	1999	w:AZ	0–4	100*	*
MD: McCasson	1999	w:IA	32	*	*

*denotes fewer than 5 tokens in the cell.

In contrast, in the academic informative registers collected for MICASE the *not*-contraction percentages are very low; this is consistent with a hypothesis that information is critical in a classroom setting. Percentages are somewhat higher in more interactive MICASE situations, but not significantly so. Similarly, *not*-contraction is quite low in all three ATC corpora, presumably because the information carried by the negative is so critical in that register (see, for example, (3f), (4d–f), and (6a)). We conclude that, other things being equal, Aux-contraction is more common in informative situations that are unscripted, but that this effect is neutralized in the written or scripted informative prose to which we have access.

British literary dialogue. Literary use of contraction provided ample data for analysis. Here, too, if the primary influence on contraction choice were the salience of *not* we would expect narrative to favor Aux-contraction and dialogue (or at least friendly dialogue) to favor *not*-contraction. Table 5 compares the different literary corpora. In fact, for all authors shown in Figures 1 and 2, narrative tokens were frequently uncontracted, while dialogue was increasingly contracted. This fact supports the dichotomy we have proposed between informative and interactive modes of presentation. However, the evidence from dialogue tokens on Table 5 shows that, when analysis is limited to {*is not*, *are not*}, *not*-contraction in dialogue is not directly correlated simply with narrative vs. dialogue, time, or, for that matter, region.

Table 5 contrasts narrative prose (n) and dialogue (") segments from British and American literary corpora. In theory, narrative/informative prose should have lower *not*-contraction than dialogue, which is pseudo-interactive. Since contraction in narrative prose is almost nonexistent for most authors in Table 5, this hypothesis is confirmed. However, it does not tell us whether this is a useful hypothesis for understanding variation between the two contraction strategies.

The dialogue data in Table 5 was isolated from narrative data, but individual characters were not isolated from each other. This did not introduce problems for texts which did not purport to differentiate among speakers by their region. For example, Tyler's characters are from Baltimore, Cleary's are from the northwest coast, and Keillor's are from Lake Wobegon, Minnesota. The dialogue data should not be merged for authors whose characters are intended to be from different locales or social groups. Consequently, the present results for literary dialogue should be regarded as preliminary. Our intention was merely to determine whether any systematic features could be recovered from literary dialogue and whether those features reflected dialect, register, or change in time.

Figure 1 reflects the evidence for British dialogue data in Table 5. It appears to show that time is not a major factor in the analysis of the choice between the two contraction strategies used in British literature. However, we discover that regional dialect of the author, or of the purported speaker, is relevant to the choice of contraction strategy. Figure 1 appears to confirm Trudgill's general claim that the further north one goes on the British Isles (as long as one doesn't go past York), the more likely speakers are to use Aux-contraction. Emily and Charlotte Brontë and Elizabeth Gaskell were from Yorkshire, as were their characters, and

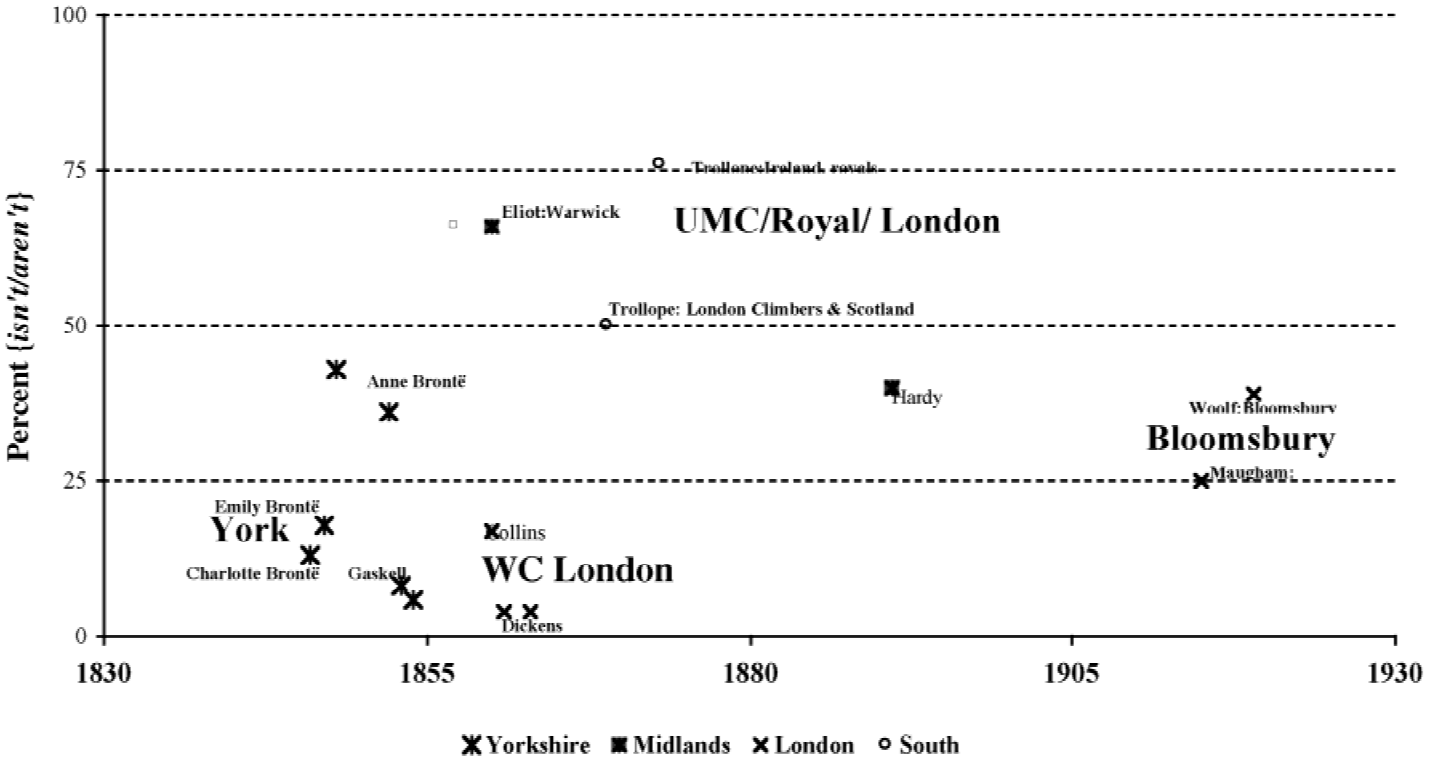

FIGURE 1. {*isn't*/*aren't*}/total contractions for dialogue in British literature.

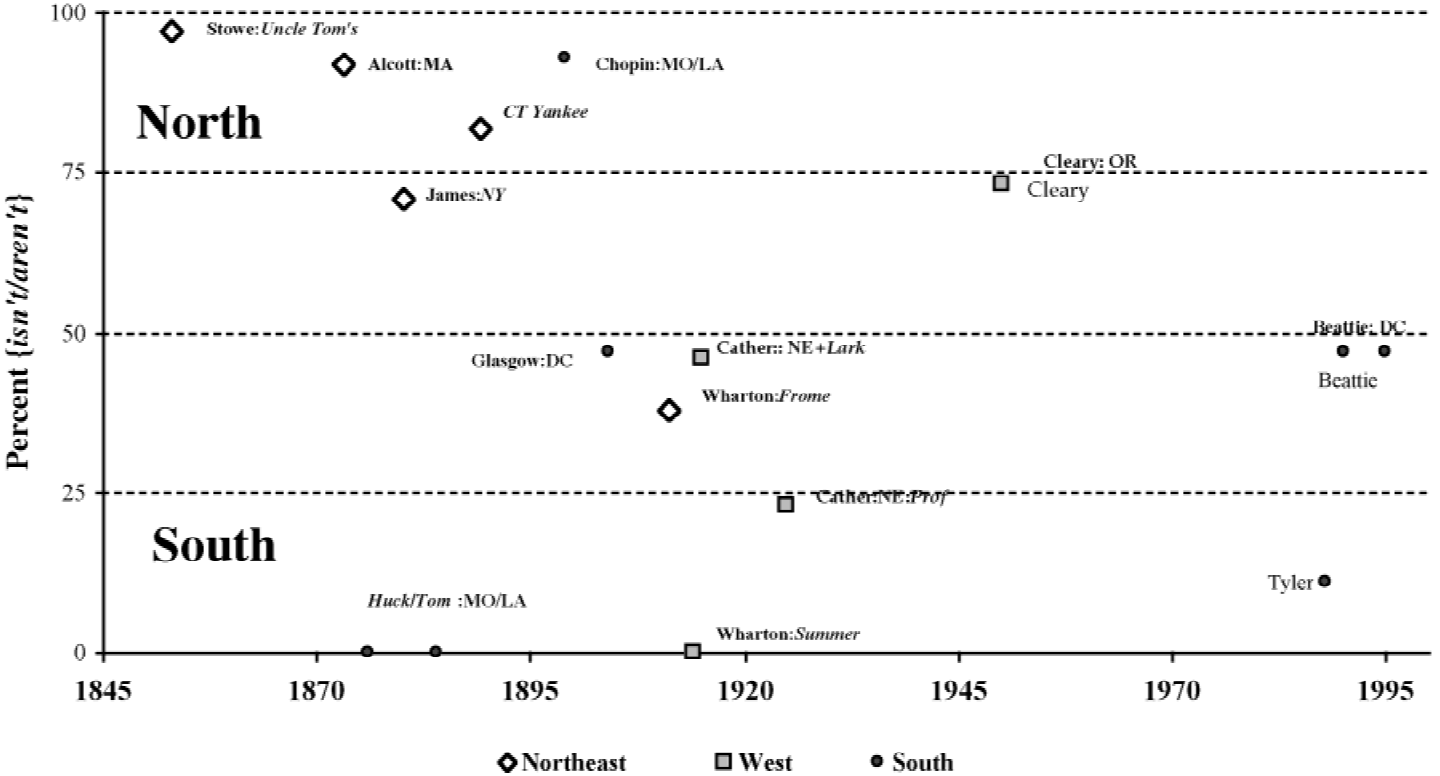

FIGURE 2. {*isn't/aren't*}/total contractions for dialogue in American literature.

their characters' *not*-contraction percentages are very low. Of the three Brontë sisters, Anne, who wrote while she was a governess further south, is the only one with *n't* over 20%; Charlotte, who is reported to have retained her dialect phonology most markedly (Barker, 1994), has the lowest percentages. In fact, the difference between Charlotte and Emily Brontë, Elizabeth Gaskell, and Anne Brontë is quite significant ($p < .001$).

Note that as one moves south the percentages rise. Eliot's characters from Warwick (*Mill*) and Hardy's Dorset characters (*Tess*) are from the rural midlands or southern counties. While Hardy's Dorset speakers (40% *n't*) do not really differ from Anne Brontë's vaguely upper crust southern speakers, Eliot's do (66% *n't*). Hardy and Anne Brontë do not differ significantly from each other, but they differ significantly from Eliot ($p < .03$), who differs even more strongly from the Yorkshire characters of Gaskell and the other Brontë sisters ($p < .0001$).

On the other hand, the Londoners certainly do not fit the pattern projected by Trudgill. There appears to be two distinct London patterns. Characters in Dickens and Collins have very low percentages, while Maugham's and Woolf's upper middle-class Bloomsbury characters have higher percentages ($p < .01$). On the other hand, even the middle-class speakers of Woolf and Maugham have significantly lower percentages than Trollope's and Eliot's speakers ($p < .0001$).

American literary dialogue.[20] The authors from northern United States in Table 5, whose results are plotted in Figure 2, have higher percentages of *not*-contraction than the southern authors, and the difference is significant ($p < .0001$). Stowe (from Massachusetts), Alcott (from Massachusetts), and Cleary (from Oregon) use higher percentages of *n't* contraction than even Trollope. Note that, while Stowe may have thought she was writing southern dialogue, the data indicate that her ear was from Massachusetts. Note also that Cleary, who is from the northwest, has contraction percentages more like the northeast than the midwest.

James' (aristocratic New York) *not*-contraction percentages are almost as high as the New Englanders', but are much higher than Wharton's. However, while both Wharton and James were from a New York Brahmin background, he was portraying members of their class, while her characters were rural and from other parts of the country. Note that Chopin (from northern Virginia) and Ann Beattie (from the Washington, DC area) have similar percentages, although they are a century apart. In contrast, although all Tyler's characters are from Baltimore, they have much lower percentages, perhaps because she herself is from the south. Authors from the midlands or the west like Cather (Nebraska) have lower percentages, but not as low as southerners like Tyler (North Carolina) or Twain's Mississippi delta residents.

The results also demonstrate that some individual authors are capable of varying contraction strategies.[21] Mark Twain has southern speakers like Huck Finn, Tom Sawyer, and their contemporaries using Aux-contraction or *ain't*, while the Connecticut Yankee and King Arthur's courtiers use *not*-contraction more consistently than either Wharton's Yankees or Trollope's courtiers. The difference

between Twain's Huck and friends and the Yankee and courtiers is significant ($p < .002$). The dialogue tokens for Trollope actually include Scottish and Irish characters along with standard, royal, and a broad range of classes for southern British characters.

Wharton differentiates between relatively urban and rural speakers: rural speakers are more likely to use Aux-contraction. Ethan Frome and his New England neighbors use *n't* fairly consistently, although not as consistently as the narrator, while the would-be midwesterners (in *Summer*) use none, and the difference is significant ($p < .04$). While this may not be an accurate rendition of small town Illinois contraction strategies at the turn of the last century, it certainly must reflect Wharton's perception that people from small towns in the west use Aux-contraction more consistently than do those from New England and Twain's perception that one of the things laughably different about Yankees and high fallutin' Brits is their greater tendency to use {*isn't, aren't*}.

Cather also varies her speakers' contraction strategies depending on their location. Although she moved to Nebraska at age eight, she moved to the east 20 years before she published *The Song of the Lark*, whose characters escape from Nebraska and use *not*-contracted tokens approximately half the time. *The Professor's House* was written ten years later, but in that work the characters are locked in Nebraska. The dialogue reflects her perception that east coasters (and middle-class/literary speakers) use *not*-contraction more than rural prairie speakers.

It seems clear that writers attempt to portray the contraction strategies they understand to be appropriate for the region and social group their characters are from. Whether they are conscious of this variable is another matter. One salient influence on {*is not, are not*} contraction strategies in both British and American dialects is region; social class is very likely to be an important factor, both in London (Dickens and Collins vs. Woolf and Maugham, $p < .01$) and in the southern United States (Chopin vs. Twain, $p < .0001$).

From this subcorpus we conclude that, when scripted speech is intended to reflect an interactive register, *not*-contraction is not as high as we found it to be in scripted informative data, and that authors' contraction choice appears to reflect their perceptions of regional differences. This is true both for American and British authors.

Use of contraction in actual interactions

British interactive conversations. We have carried out a pilot study of the COLT data (London), while Tagliamonte and Smith (in press) analyzed data from older rural interviewees and Gasparrini (2001) analyzed data from the BNC—sportscasts, club meetings, call-in and chat programs (cf. Hutchby, 1996, 1999)—which she characterized as at least 70% adversarial (personal communication, Nov. 2001).[22] Table 6 compares these British interactive corpora. Note that Tagliamonte's register is probably supportive, while Gasparrini's is perhaps somewhat less adversarial than political debates. To the degree that these

data can be said to reflect register variation, the results appear to support our hypothesis. Gasparrini's adversarial data have lower *not*-contraction percentages than Tagliamonte and Smith's interviews from the same region.

These percentages can be usefully compared with the literary results shown in Table 5, although interpreting the results is sometimes difficult. Our analysis of the COLT data shows that very few *not*-contractions occur in friendly conversations among London urban youth and their adult interlocutors. Although these data were gathered in the 1990s, they are consistent with the pattern reflected in Dickens' and Collins' dialogue. Tagliamonte and Smith's Devon results are fairly consistent with Eliot's and Hardy's rural characters' speech. On the other hand, *not*-contraction is far more common for Tagliamonte and Smith's 20th-century speakers than for the Brontës' and Gaskell's 19th-century dialogue. Perhaps this is due to the changing demographics of the area. Tagliamonte and Smith's Irish and Scottish rural speakers have very few *not*-contractions. This appears to contrast with Trollope's characters, since Scottish (in *Eustace Diamonds*) and Irish characters (in *Phineas Finn*) are included with the parliamentarians and royals. Clearly, further studies of contraction should isolate characters from different backgrounds.

American interactive conversations. Table 7 compares the United States supportive interactive corpora. Note that the British dialects have a broader range of *not*-contraction than American dialects, but there is still a wide range of *not*-contraction percentages available to these speakers.

The SWB conversations demonstrate that, even for urban upper middle-class employees of the technological revolution conversing on hyperpolite topics, speakers from the north were significantly more likely to use *not*-contraction than were those from the south (Deckert & Yaeger-Dror, 2000). The results for all supportive corpora in Table 7 are consistent with those of the SWB corpus. CHILDES (Pittsburgh), Upholstery Shop (New York City), and Segrin (Kansas) speakers did not differ significantly from the SWB speakers from their area. The results, like the literary results, are consistent with Feagin's (1979) surmise that the percentage of Aux-contraction from middle-class Anniston (Georgia) residents was higher than the percentages that would be found further north.

Compare the *not*-contraction percentages in dialogue with those found in the conversational corpora. SWB's western speakers (apparently from Texas) have low *not*-contraction percentages, while Cleary (Oregon) and Nixon (California) have high percentages.

Recall, as well, that American dialogue (in Table 5) reflects both a north–south distinction and an urban–rural distinction, which cannot be tested using the conversational data. Neither Cather's nor Wharton's rural characters reflect the late 20th-century results for middle-class urban conversationalists, nor is there a corpus of rural conversations similar to Tagliamonte and Smith's against which these literary authors' intuitions can be judged.

Unfortunately, the SWB corpus cannot be used to analyze speaker stance since there are so few remedial tokens that it precludes an analysis of stance (Yaeger-Dror & Hall-Lew, 2000). Moreover, while speakers were coded for where they

grew up until age 12, they were almost all Texas Instruments employees, and all appeared to be middle-class. Although the SWB data permits the analysis of the influence of region on contraction, neither social class nor interactive stance can be analyzed using this corpus. While statistical analysis of stance and prosody are still in progress for the Segrin corpus, it appears that Segrin's couples do not alter their contraction strategy relative to the task that they are performing, since both the supportive and remedial task settings have identical percentages, which may actually reflect the fact that the task is, for these couples, both supportive and informative.

The only corpus that differs radically from our expectations based on the SWB data is the Tripp/Lewinsky corpus. Since one participant is from the west coast and the other is from the greater New York City area, one would project that at least Tripp would have higher *not*-contraction. Although a close analysis of the stances taken in those conversations is beyond the bounds of the present study, we project that it would provide data for an in-depth analysis of stance.

Table 8 considers contraction use by adversarial speakers. Debaters clearly use full form more than other speakers do. In fact, Kennedy and Nixon, with their heavily scripted debate, use full form almost as much as the written informative texts. These high percentages are consistent with an understanding that the information carried by *not* is high, and that the debaters are ratified to express negation openly, both because the negative carries important information and because there is a preference for disagreement in this register.[23] Following this same logic, debaters should also use Aux-contraction more than *not*-contraction so as to permit emphasis on the negative consistently. However, Table 8 reveals that *not*-contraction is not as uniformly low in presidential debates as our initial hypothesis projected. Comparing Tables 7 and 8 it is clear that the debaters' dialect area—rather than their position along a time-line or their adversarial stance—appears to be the most salient factor influencing contraction strategy. Southern candidates like Carter, Perot, Gore, and Clinton favor Aux-contraction, not just for {*is not/are not*} but for {*have not*} as well, while candidates from the north favor *not*-contraction. In fact, most northern speakers are 88 times as likely to use *not*-contraction as the southern politicians ($p < .0001$). Three speakers do not fit this regional pattern. Both Bush family members use *not*-contraction considerably less than their Ivy League background would make appropriate ($p < .0001$) but considerably more than their Texan pretensions would project ($p < .009$). Kennedy's percentages are neither as high as his New England background would project, nor as low as the Kennedy Boston-Irish label would project.

One of the advantages of using the presidential data is that we can compare a given speaker using two different registers. In the present instance, we can compare these speakers' adversarial stance (Political Debates in Table 8) with their usage when answering questions during a news broadcast—an informative stance (Q/A in Table 4.) One would project that in a debate a speaker would avoid reducing the negative both to express himself clearly and to maximize the face threat to his opponent. In a Q/A session, a president would want to avoid reducing negatives primarily to express himself clearly. One might

then hypothesize that in the latter the {*isn't/aren't*} percentages would be higher than in the former. A comparison of the relevant information in Tables 4 and 8 supports that hypothesis. Note that the percentages for each speaker are fairly consistent. The southern politicians (Carter, Clinton, Perot, Gore) have very low percentages in both registers, and the northern speakers (Ford, Kennedy, Reagan) have higher percentages in both registers. Speakers from the west (Nixon, Mecham) also have higher *not*-contraction percentages. All other things being equal, the Republicans tend to have higher percentages than the Democrats, and this holds for both registers. However, register also influences the results and does so in the direction that we hypothesized. All presidents except for Reagan have lower percentages of {*isn't aren't*} in the debates than in the Q/A sessions (with the difference significant for Kennedy, Nixon, Reagan, and Bush individually, as well as for the combined results of all presidents except Reagan; $p < .0001$).

CONCLUSION

What have we learned from this study about English negative contraction? Given that previous studies have already documented a broad range of variation in contraction strategies, and the fact that morphology, syntax, dialect, and register all play a significant role in that variation, it is now clear that both linguistic and demographic factors should be considered in any analysis of contraction strategies. It is also clear that one should not merge contraction data from nonhomogeneous groups of speakers (even in one text) or from different registers (even when used by a single speaker).

Analysis of data from a large corpus can often uncover linguistic patterns that contradict our expectations of how our language is used; in fact, this analysis developed out of just such a contradiction. In the course of a qualitative analysis of the variation in disagreement strategies, Yaeger-Dror (1997) assumed—along with English as a second language teachers, register theorists, and speech scientists—that American English speakers would use *not*-contraction almost categorically, and that *not*-contraction would be consistent across all verb+*not* locutions; to the extent that Aux-contraction might be used, she assumed that it would reflect only the register-driven need to focus on or reduce the negative. However, in the process of analyzing disagreement strategies based on a subset of the SWB corpus, it became clear that these initial assumptions were not supported. It became obvious that region was a major influence on contraction choice for *is not* and *are not*. Expanding the corpus to include other registers appears to support the hypothesis formed from the SWB results. This theory is supported by data from England as well as the United States and Canada, and it is supported by literary dialogue as well as different conversational registers.

The analysis was carried out on the comparison of *not*-contracted and Aux-contracted forms of {*is not, are not, have not, will not, has not*}. Preliminary analysis showed that in American English declarative sentences Aux-contraction was interesting primarily for {*is not, are not*}, and in fact, our intuitions about con-

traction strategies for *is not* and *are not* were supported by fewer than half of the tokens for most of the speakers analyzed. Despite the intuition that we use *isn't* and *aren't*, despite the fact that *not*-contraction is the pattern used overwhelmingly in all other American negative contractions, and despite the fact that *not*-contraction would more obviously support the Social Agreement Principle (Schegloff, Jefferson, & Sacks, 1977; Yaeger-Dror, 1997), the evidence shows that dialect area is a significant factor influencing *not*-contraction, with *not*-contraction preferred in the north (including New England, the northern cities area, the Ontario area, and the northwest) and Aux-contraction favored in the southeast and southwest. We also found significant regional variation in the British data.

For *is not* and *are not* the choice of one contraction strategy over another appears to be significantly correlated with the speaker's dialect area in conversational, adversarial, and literary situations. Feagin, as a native speaker of inland southern dialect, assumed that other dialects use Aux-contraction less consistently than her Anniston speakers (1979), but she had no data from other dialect areas with which to make a systematic comparison. The data from the SWB corpus document the fact that, when we have comparable conversational data from middle-class speakers whose speech is included in computerized megacorpora, speakers from distinct regions have statistically different tendencies to prefer one contraction strategy over another, with southern speakers using Aux-contraction more than northern speakers. The more the corpus is expanded to include speakers from different regions, the more diverse the regional pattern becomes.

The fact that *not*-contraction is significantly more common in the speech of southern middle-class Americans would never have been considered without the evidence from an analysis of the SWB data. Introducing data from debates and literary texts demonstrated that region provides more flamboyant contrasts than register, but that some register distinctions can also be distinguished. The introduction of political and literary data permits more time depth to be added to this analysis; while an earlier study found that the contraction of *not* in the "other" verbs (in both these literary and political corpora) is a change in real time (Yaeger-Dror et al., in press), region appears to be the primary influence on the choice of *not*-contracted form over Aux-contracted form. None of the United States corpora reveal a major change in time; only the comparison of 19th-century York dialogue with Tagliamonte and Smith's interviews leads us to infer that a change may be taking place in British dialects. The ideal, of course, would be to be able to code a single corpus for dialect, register, and stance in order to compare the relative importance of these factors simultaneously. Acquiring even relatively comparable data sets which permit dialect and register comparisons has not been simple; of the corpora we have analyzed, only the contrast between presidential registers permits such a systematic comparison.

Difficulties comparing corpora

We found that neither register nor dialect can be teased out of the news files available from LDC or Oxford. Because journalism still retains full form at least half the time, a very large corpus would be necessary to permit a useful analysis.

While the collection of a dialect-coded journalism corpus was so time consuming that the data gathered to date had too few contractions to be useful for the analysis of these two contraction strategies, a great deal was learned by trying to collect data that would be sensitive to region. Even looking at the most informative journalistic register, science or health news, the researcher must be especially careful to check that the copy was written by a local reporter rather than just cobbled together from the wire services. Book reviews or articles with first person narration cannot be compared with other reviews. Reviews or articles whose authors take an adversarial stance cannot be compared with those that are not adversarial. Maintaining these "rules" is always possible, but is certainly more time consuming. On the other hand, most hometown newspapers can now be accessed over the web by a researcher interested in such research, and all journalists questioned to date have been very responsive to the researcher interested in dialect information.

Even the analysis of literary prose was complicated by availability of appropriate comparable texts from different generations of speakers. This preliminary study determined that scanning from books is a prohibitively wasteful use of time, but many text files are available online for a reasonable sum. In the process of picking text files to download, it became obvious that good literature is available online for older sources, while horror, mysteries, and sci fi are available for the late 20th century, and so change in contraction preferences is not simple to judge.

In any study of literary texts, care must be taken to compare only similar genres. However, when comparisons are possible, they reveal a robust dialect influence on {*is not*/*are not*} contraction in dialogue. Comparison of authors from the same region (Dickens and Collins vs. Maugham and Woolf; Twain vs. Glasgow or Chopin) reveals the importance of including class as a variable as well.

Looking specifically at dialogue, while British authors like Trollope and Eliot are very likely to attribute *not*-contraction to the class of the speaker, the American authors appear more attuned to a character's dialect area. Mark Twain, Willa Cather, and Edith Wharton clearly perceived that speakers from different areas must have different contraction strategies. Since the variation within dialogue may reflect the authors' stereotyped intuitions rather than an accurate portrayal of a specific regional or class dialect, the simplest analyses can be designed around authors whose characters all share the same regional and class characteristics as the author. The larger and more varied the cast of characters, the more complex—but also the more necessary—the concordance should be. To the degree that we are interested in the authors' linguistic intuitions and stereotypes, characters in the more socially complex writings should have separate concordances. Note also that the preliminary results reported here appear to support the claim that regional differences in contraction strategies may be so robust that neither time nor genre appears to have an impact which can obliterate their influence. Higher *not*-contractions occur both in the mid-19th-century New England and British southern middle-class dialogue and in the late 20th-century conversations of people from those areas. Lowest *not*-contractions occur for working-class Londoners today, in 1850s dialogue, and for southern United States speakers, whether they are debaters, SWB phone callers, or Huck Finn.

Considering only the audio corpora from the United States, the situation is likewise both heartening and disheartening. Whereas ten years ago the audio corpora available had been collected by engineers and mostly consisted of single words spoken by engineers or news re-read by professionals, now there are tapes of actual telephone conversations (from LDC) and classroom interactions (by MICASE). Unfortunately, even today most researchers assume that all academics and other middle-class Americans speak an American *Koiné* (at least syntactically), and so dialect information is generally not retained and can never be accessed later. While the SWB corpus is far from ideal in its register design (Yaeger-Dror & Hall-Lew, 2000), analysis of this corpus certainly demonstrates that a speaker's dialect area cannot be ignored, even for Texas Instruments employees using their most supportive style for talking to strangers about topics transparently devised for the collection of speech data. We hope that data from a range of registers with dialect information tabulated will soon be available, and we should take advantage of it.

The results of the present analysis can be used in many ways; a few of these follow. Earlier we noted that dialect has an even greater impact on contraction strategies in imperative and interrogative sentences. Analysis of this variation has yet to be attempted.

Both authors' intuitions and Feagin's rural data lead us to believe that rural speakers speak differently from city residents in the same region. Tagliamonte and Smith have collected data to test that hypothesis in England, and Feagin has collected appropriate data for the region around Anniston, but rural data should be considered more systematically in our studies of dialect variation.

While hunting for perfectly comparable and perfectly socially situated corpora, this analysis has provided evidence that a factor that has not been analyzed extensively in the United States and Canada must be included as a significant factor group in any analysis of contraction strategies. We must consider each verb and each dialect region separately, along with the continuum from informative to interactive register and the continuum from supportive to adversarial social stance. These insights are now being incorporated into an ongoing analysis of the interaction between linguistic, pragmatic, and prosodic parameters in the study of *not*-negation.

During the analysis of the sound files, we found that speakers who reported themselves as New Yorkers but were calling from Texas used neither New Yorker phonology nor a unified contraction strategy, but their phonology was clearly Texan. We conclude that any dialect-relevant analysis should supplement self-report information with sociophonetic dialect coding drawn from Labov's Telsur project (www.ling.upenn.edu/phono-atlas/home.html) or other recent dialect studies. This task should be carried out by sociolinguists; it will be useful for speech engineers as well as for sociolinguists, since speech scientists are constantly seeking information that can improve their recognition algorithms, and improved dialect monitoring will help them achieve their goals.

Sociolinguists should explain to the organizations that have funding to gather speech data (like LDC) how important accurate dialect information is for a meaningful study of syntax. We should also clarify to other researchers that both dialect and register information should be carefully monitored even for

studies of variables that appear to be independent of one or both of these factors. An ideal corpus would provide self-report information on both pre- and post-adolescent residence, as well as carefully chosen phonetic information and accurate information about the register and stance of the recording.

While we have already documented that the United States corpora need to be coded for dialect more carefully, it is also important to note that the British corpora could be coded more carefully for register and stance. An accurate evaluation of the importance of register and stance will be much easier when corpora are coded for these variables as carefully as the British corpora are coded for dialect.

Social psychology of language change can also use the corpora that were used in this study. Those whose phonology does not match their purported dialect area may still have syntax that is consistent with where they grew up, or vice versa. The limited amount of information available from this study (from the Bushes and from the SWB New Yorkers) would imply that both phonology and syntax can vary below the speaker's awareness. However, specific studies should be designed to test this theory systematically.

The presidential libraries have a plethora of data that could be used for analysis of stance taking and variation in both stance and social psychological variables. This study has confirmed that contraction is unconsciously used strategically to display a specific persona to the listening audience; any studies of politicians' speech could be expanded to permit analysis of the social psychology of the politicians' accommodative tendencies.

As already suggested by Tagliamonte and Smith (in press), we can integrate information from their 20th-century British results with the information about United States and Canadian settlement patterns to gain a better understanding of the types of changes that took place as this continent was settled.

We can also explore why {*is not*, *are not*} are the only holdouts from *not*-contraction in American English. Just as Phillips (1984) used the simplification of the (yu) glide to explore the relationship between word frequency, speakers' social status, and resistance to and rate of change, a similar study would probably provide interesting sociolinguistic conclusions for the analysis of contraction and in turn for our understanding of language change phenomena in general.

This study has provided a pilot for collection and analysis strategies for large corpora. We hope that future studies will benefit from our mistakes and take advantage of our conclusions.

NOTES

1. We are grateful to an anonymous reviewer for reminding us to emphasize that editorial style regulations of different publication houses provide a conservative bias in published materials. Thus, this shift can be regarded as a change in genre-based expectations, which is independent of the amount of contraction that may or may not have occurred in conversation at a given point in the past.

2. Unfortunately, certain distinctions are conflated by the decision to merge editorials with letters to the editor and sports and social reporting with international news.

3. Cross-cultural differences have also been the focus of much recent work, showing, for example, that girls are more adversarial in play groups in China than in the United States (Kyratzis & Guo, 2001), that Greeks (Kakava, 2002) and Ashkenazim (Schiffrin, 1984; Tannen, 1984) do conver-

sations more adversarially than Americans, that British and United States debaters are more adversarial than French debaters (Yaeger-Dror, 2002b), and that legal interactions are more adversarial in the United States than in Britain (Kurzon, 2001). Such cultural variation will not be discussed here.

4. We are most grateful to Tagliamonte for prepublication access to these results.

5. The codes for different corpora can be found in the Appendix. Following the code, the speaker and record number are listed.

6. The fact that the choice of *'ll not* or *won't* is correlated with dialect area is clearly relevant to a larger study; in addition, given that *won't*, like *ain't*, is clearly heard to be negative even before *n't*, an adequate register-relevant analysis of *'ll not/won't* would have to take that fact into consideration.

7. Tokens cited were gathered from the corpora discussed later, with the number indicating the line of text as defined by the developer of the cited corpus.

8. Journalists' bios are generally available online, and where they are not or where sufficient dialect-background information is not included, the individual journalists have shown themselves to be pleased to be asked where they are from, and they will give detailed information on where they went to school and what their family background is. Journalists who responded also maintain that editors do not edit their choice of contraction strategy. Similarly, biographical information for specific authors is generally available online for literary or informative writing being downloaded; often supplementary information, unavailable to your reference librarian a few years ago, can now be found with the help of google.com.

9. The Strathy collection of Canadian news writing should provide the advantages of a large corpus without the disadvantages, but access to that corpus is severely restricted.

10. A number of the Canadian reviews discussed books about Quebec separatism or the biographies of Quebec politicians. These reviews were written by the political correspondent rather than the book editor, and since they were quite adversarial in stance and none of the northeastern United States book reviews were adversarial, there was no possible comparison; adversarial reviews were discarded.

11. Initially, the only way to incorporate recent texts into an analysis is to scan them to permit a concordance to be run. Unfortunately, this is such a slow process that almost any e-book price on the market will soon be cheaper than sending a student to scan a segment of a book. We learned this the hard way.

12. On the other hand, the smart-mouthed genre of kid lit known and loved in recent years differs radically from the genre of kid lit available from earlier generations. Neither *Little Women* nor *Pollyanna* (which are available) is directly comparable to Cleary's work. If equivalents do exist, they are certainly not available on the web, and so time depth is not available for this genre.

13. We assume this reflects the fact that neither pilots nor traffic controllers are necessarily local to the airports where their speech was recorded. Obviously, information-processing is far stronger in ATC than in other registers studied.

14. Regional information for presidential aspirants can be a mixed blessing. Politicians are talking for both the immediate and listening audiences (Bell, 1984), and so accommodative tendencies (Giles & Coupland, 1991) can be doubly difficult to determine. To cite only two examples: in the present corpus, Bush-*père* (csdl.tamu.edu/bushlib/) is clearly a product of the northeast, but in his 1992 election bid to white-wash the media's characterization of him as a wimpy, preppy member of the eastern establishment, he converged toward the social psychologically more macho region where he had settled by adopting a few Texas dialect features. For example, one form of convergence introduced tokens of *wuhdn't* and *idn't*.

In campaign ads broadcast within the southwest, Bush-*fils* tells us to take care of our *chirren*, although he also has a prep school/Ivy League education, making it unlikely that at home he uses the dialect persona adopted for these ads. One wonders whether he was coached for alternative dialect profiling for ads to be broadcast outside the south and southwest; certainly even the comedian who has been hired to parody him finds his complex pattern of accommodation too tough an act to follow (Leland, 2001).

In short, while this study has taken advantage of an expanded array of textual and audio data from the presidential archives, we are all aware that any dialect information we infer from political debates, where the public is an acknowledged overhearer of the proceedings (Bell, 1984), may well be tainted by conscious accommodative choices the speakers have made relative to their rival candidates or relative to the image they wish to project to their home constituency and/or the larger national audience.

15. See also Pérez de Ayala's (2001) discussion of the Hansard texts for the British Parliament.

16. For the most part they had been married within the last two years and were still together six months later.

17. Gasparrini (2001), however, in her study of a subcorpus from the BNC did find that the distinction was useful. In her subcorpus the auxiliary was somewhat more likely to be Aux-contracted.
18. Comparisons are for *isn't*/{*isn't*+*'s not*} and *aren't*/{*aren't*+*'re not*}.
19. While intuitions are unreliable, we find *It's not* . . . as acceptable as *She's not*. Kjellmer (1998) confirmed this intuition by showing that *it* and *that* permit Aux-contraction at twice the rate of *s*/*he* or *they*. *Jane's not* . . . sounds better than *Tommy's not* . . . (So perhaps a preceding consonant in a stressed syllable favors Aux-contraction more than an unstressed vowel?) Moreover, *It's really not* . . . is more acceptable than *It really isn't* . . . (not to mention the form which would provide a preceding vowel: *it really's not* . . .). None of the studies presently available permit an adequate comparison of the data. Since preceding stress placement appears to us to be more important than initially assumed, study of prosodic, phonological, and syntactic context is incorporated into the ongoing analysis of prosodic information and *not* prominence and will be reported in subsequent work.
20. It is still impossible to acquire Canadian fictional data except for exorbitant fees.
21. Further evidence can be found in Kretzschmar (2001).
22. Again, we are grateful to Tagliamonte and Smith and Gasparrini for prepublication access to their results, which have beefed-up the cross-register comparison of British data.
23. See Yaeger-Dror (2002a, 2002b) for discussion of a contrasting pattern in French debates, as well as for evidence that prosodic prominence can and does take over in adversarial interactions.

REFERENCES

Al-Khatib, Mahmoud. (1997). Provoking arguments for provoking laughter: A case study of the candid camera TV show. *Text* 17:263–299.

Aston, Guy, & Burnard, Lou. (1998). *The BNC handbook*. Edinburgh: Edinburgh University Press.

Barker, Julia. (1994). *The Brontës*. New York: St. Martins.

Bell, Allen. (1984). Language styles as audience design. *Language in Society* 13:145–204.

Biber, Douglas. (1988). *Variation across speech and writing*. Cambridge: Cambridge University Press.

Biber, Douglas, Conrad, Susan, & Reppen, Randi. (1998). *Corpus linguistics: Investigating language structure and use*. Cambridge: Cambridge University Press.

Biber, Douglas, Finegan, Edward, & Atkinson, D. (1993). ARCHER and its challenges: Compiling and exploring a representative corpus of historical English registers. In U. Fries, G. Tottie, & P. Schneider (eds.), *Creating and using English language corpora: Papers from the Fourteenth International Conference on English Language Research on Computerized Corpora*. Amsterdam: Rodopi. 1–13.

Biber, Douglas, Johansson, Stig, Leech, G., Conrad, S., & Finegan, E. (eds.). (2000). *Longman grammar of spoken and written English*. London: Longman.

Blum-Kulka, Shoshana, Blondheim, Menahem, & Hacohen, Gonen. (2002). Traditions of dispute: From negotiations of Talmudic texts to the arena of political discourse in the media. *Journal of Pragmatics* 34.

Bresnan, Joan. (2000). Explaining morphosyntactic competition. In M. Baltin & C. Collins (eds.), *Handbook of contemporary syntactic theory*. Oxford: Blackwell.

Clayman, Steven. (2002). Disagreements and third parties: Dilemmas of neutralism in panel news interviews. *Journal of Pragmatics* 34.

_____ (in press). The news interview. In Neil Smelser & Paul Baltes (eds.), *International encyclopedia of behavioral sciences*. Oxford: Elsevier Science.

Cleary, Beverly. (1968). *Ramona the pest*. New York: Scholastic Books.

Corsaro, William, & Rizzo, Thomas. (1990). Disputes in the peer culture of American and Italian nursery-school children. In Allen Grimshaw (ed.), *Conflict talk*. Cambridge: Cambridge University Press.

Deckert, Sharon & Yaeger-Dror, M. (2000). Disagreement, contraction and dialect: Evidence from a large corpus of American English. *CLIC* 2:49–59.

Denison, David. (1999). Syntax. In Suzanne Romaine (ed.), *Cambridge history of the English language* (Vol. 4). Cambridge: Cambridge University Press.

Dumas, Bethany. (2001). *Network English: Fact or fantasy?* Paper delivered at ADS session 22, LSA, Washington, DC.

Feagin, Crawford. (1979). *Variation and change in Alabama English: A sociolinguistic study of the White community.* Washington, DC: Georgetown University Press.

Flora, J., & Segrin, Chris. (2000). Affect and behavioral involvement in spousal complaints and compliments. *Journal of Family Psychology* 14:641–657.

Gasparrini, Désirée. (2001). *It isn't, it is not* or *it's not? Regional differences in contraction in spoken British English*. Master's thesis, University of Zurich.
Giles, H., & Coupland, N. (1991). *Language: Contexts and consequences*. Pacific Grove, CA: Brooks/Cole.
Goffman, Erving. (1971). *Relations in public*. New York: Harper.
_____ (1981). *Forms of talk*. Oxford: Blackwell.
Goodwin, Marjorie. (1983). Aggravated correction and disagreement in children's conversation. *Journal of Pragmatics* 7:657–677.
Goodwin, Marjorie, Goodwin, Charles, & Yaeger-Dror, Malcah (2002). Multi-modality in girls' game disputes. *Journal of Pragmatics* 34.
Hazen, Kirk. (1996). Linguistic preferences and prescriptive dictum: On the phonological and morphological justification of *ain't*. In J. Arnold, R. Blake, B. Davidson, S. Schwenter, & J. Solomon (eds.), *Sociolinguistic variation: Data, theory and analysis*. Stanford: CSLI. 101–112.
Heritage, John. (2002). The limits of questioning: Negative interrogatives and hostile question content. *Journal of Pragmatics* 34.
Hiller, Ulrich. (1987). An investigation into the choice between the two contracted variants of negated English auxiliaries. *Die Neuren Sprachen* 86:531–553.
Hirschberg, Julia. (1993). Pitch accent in context. *Artificial Intelligence* 63:305–340.
Hoyle, Susan, & Adger, Carolyn Temple (Eds.). (1998). *Kids' talk: Strategic language use in later childhood*. New York: Oxford University Press.
Hudson, Richard. (2000). I amn't. *Language* 76: 297–323.
Hutchby, Ian. (1996). *Confrontation talk*. Mahwah, NJ: Erlbaum.
_____ (1999). Rhetorical strategies in audience participation debates on radio and tv. *Research on Language and Social Interaction* 32:243–267.
_____ (2001). 'Oh', irony and sequential ambiguity in arguments. *Discourse and Society* 12:123–141.
Ilie, Cornelia. (1999). Question-response argumentation in talk shows. *Journal of Pragmatics* 31:975–999.
Jacobs, Scott. (2002). Maintaining neutrality in third-party dispute mediation. *Journal of Pragmatics* 34.
Jefferson, Gail. (2002). Is *no* an acknowledgment token? *Journal of Pragmatics* 34.
Jespersen, Otto. (1917). *Negation in English and other languages*. Copenhagen: A. Høst.
Johansson, S., & Oksefjell, S. (eds). (1998). *Corpora and cross-linguistic research* (vol. 24). Amsterdam: Rodopi.
Kakava, Christina. (2002). Opposition in modern Greek discourse. *Journal of Pragmatics* 34.
Keillor, Garrison. (1985). *Lake Wobegon days*. New York: Penguin.
Kennedy, Graeme. (1998). *Corpus linguistics*. London: Longman.
Kjellmer, Gören. (1998). On contraction in Modern English. *Studia Neophilologica* 69:155–186.
Kretzschmar, W. (ed.). (2001). Special issue of *Language and Literature* 10(2).
Kurzon, D. (2001). The politeness of judges: American and English judicial behavior. *Journal of Pragmatics* 33:61–85.
Kyratzis, Amy, & Giansheng Guo. (2001). Preschool girls' and boys' verbal conflict strategies in the US and China. *Research on Language and Social Interaction* 34:45–74.
Leland, John. (2001). Parody politics: Questions for Timothy Bottoms. *New York Times Sunday Magazine* 3.25.01: 23.
Lightfoot, David. (1999). *The development of language: Acquisition, change and evolution*. Oxford: Blackwell.
McElhinny, Bonnie. (1993). Copula and auxiliary contraction in speech of white Americans. *American Speech* 68:371–99.
Pérez de Ayala, Soledad. (2001). FTAs and Erskine May: Conflicting needs?—Politeness in question time. *Journal of Pragmatics* 33:143–169.
Pomerantz, Anita. (1984). Agreeing and disagreeing with assessments. In John Atkinson& John Heritage (eds.), *Structures of social action: Studies in conversation analysis*. Cambridge: Cambridge University Press. 57–101.
Phillips, Betty. (1984). Word frequency and actuation of sound change. *Language* 62:320–342.
Rand, David. (1997). A short text document for demonstrating Concorder Version 3.
Reeves, Richard. (1993). *President Kennedy: Profile of power*. New York: Simon and Schuster.
_____ (2001). *Nixon: Alone in the White House*. New York: Simon and Schuster.
Rissanen, Matti. (1999). On the order of post-verbal subject and the negative particle in the history of English. In Ingrid Tieken-Boon van Ostade, Gunnel Tottie, & Wim van der Wurff (eds.), *Negation in the history of English*. Berlin: Mouton de Gruyter. 189–205.
Sacks, Harvey. (1992). *Harvey Sacks' lectures on conversation*. Oxford: Blackwell.

Schegloff, Emanuel A., Jefferson, G., & Sacks, H. (1977). Preference for self-correction in the organization of repair in conversation. *Language* 53:361–382.
Schiffrin, Deborah. (1984). Jewish argument as sociability. *Language in Society* 13:311–335.
Scott, Suzanne. (1998). *Patterns of language use in disagreements and conflicts*. Doctoral dissertation, NAU.
Sheldon, Amy. (1996). You can be the baby brother but you aren't born yet. *Research on Language and Social Interaction* 29:57–80.
_____ (1998). Talking power: Girls, gender enculturation and discourse. In Ruth Wodak (ed.), *Gender and discourse*. London: Sage.
Stern, Sheldon. (2000a). Source material: The 1997 published transcripts of the JFK Cuban missile crisis tapes: Too good to be true? *Presidential Studies Quarterly* 30:586ff.
_____ (2000b). Letter to the Editor. *Presidential Studies Quarterly*.
Tagliamonte, Sali, & Smith, Jennifer. (in press). "Either it isn't or it's not": NEG/AUX contraction in British dialects. *English World Wide*.
Tannen, Deborah. (1984). *Conversational style*. Norwood, NJ: Ablex.
Tottie, Gunnel. (1991). *Negation in English speech and writing*. San Diego: Academic.
Trudgill, Peter. (1978). *Sociolinguistic patterns in British English*. London: Edward Arnold.
_____ (1990). *The dialects of English*. Oxford: Blackwell.
_____ (1999). *The dialects of England*. Oxford: Blackwell.
Tyler, Anne. (1988). *Breathing lessons*. New York: Random House.
Warner, Anthony. (1993). *English auxiliaries: Structure and history*. Cambridge: Cambridge University Press.
Westergren-Axelsson, Margareta. (1998). Contraction in British newspapers in the late 20th century. *Studia Anglistica Upsaliensia* 102. Uppsala: Acta Universitatis Upsaliensis.
Wharton, Edith. (1911/1969). *Ethan Frome*. Litrex Reading Room.
_____ (1917/1998). *Summer*. Litrix Reading Room.
Yaeger-Dror, Malcah. (1985). Intonational prominence on negatives in English. *Language and Speech* 28:197–230.
_____ (1991). Linguistic evidence for social psychological attitudes. *Language and Communication* 11:309–331.
_____ (1996). Register as a variable in prosodic analysis. *Speech Communication* 19:39–60.
_____ (1997). Contraction of negatives as evidence of variation in register specific interactive rules. *Language Variation and Change* 9:1–36.
_____ (2001). Primitives for the analysis of register. In J. Rickford & P. Eckert (eds.), *Style and sociolinguistic variation*, Cambridge: Cambridge University Press.
_____ (ed.). (2002a). Disagreement strategies and negation. Special issue of *Journal of Pragmatics* 34.
_____ (2002b). Disagreement strategies and negation: An introduction. *Journal of Pragmatics* 34.
_____ (2002c). Register and prosodic variation: A cross language comparison. *Journal of Pragmatics* 34.
Yaeger-Dror, Malcah, & Hall-Lew, Lauren. (2000). Prosodic prominence on negation in various registers of US English. *Journal of the Acoustical Society of America* 108:2468.
Yaeger-Dror, Malcah, Hall-Lew, Lauren, & Deckert, Sharon. (in press). Situational variation in intonational strategies. In C. Meyer (ed.), *Corpus analysis: Language structure and language use*. Amsterdam: Rodopi.
_____ (ms.). Contractions in interrogatives and imperatives.

Abbreviation	Full Title	Register	Description	url (if available)
ATC	Air Traffic Control	Informative	LDC corpus of airport interactions	ldc.upenn.edu/Catalog/LDC94S14A. html
BNC	British National Corpus	Varied	Megacorpus housed at Oxford	info.ox.ac.uk/bnc/
BNC	British National Corpus leisure	Adversarial	Call-in programs, sportscasts, meetings	info.ox.ac.uk/bnc/
Bk Rev	*NYT*	Informative	Book reviews of informative texts	—
BUR	Boston NPR Radio News	Informative	LDC corpus reread news by Boston NPR announcers	ldc.upenn.edu
§1	*NYT*/first chapters	Lit.Informative	First chapters of informative texts, downloaded.	nytimes.com/pages/books/chapters/index.html
CHILDES	CMU children & parent corpus	Interactive	Corpus of digitized conversations, aligned with transcript	childes.psy.cmu.edu
COLT	Corpus of London Teenage English	Interactive	Megacorpus of London conversations, primarily of teenagers	helmer.hit.uib.no/colt
Gasparrini	Subset of BNC	Adversarial	Sportscasts, call-in programs, club meetings	—
G/M	*Globe & Mail*, *Macleans*, *Toronto Star*	Informative	Web-downloaded from various Toronto publication archives	globeandmail.ca; macleans.ca
LDC	Linguistics Data Consortium	Varied	Megacorpora of speech available for sale	ldc.upenn.edu
MD	Mayoral Debates	Adversarial	Tucson Mayoral Primary	npr.org
Mkt	Marketplace (also see: USC)	Informative	LDC corpus of NPR economic news program	ldc.upenn.edu/Catalog/LDC99S82.html
M/L	MacNeil/Lehrer	Adversarial	Debates held during the MacNeil-Lehrer news hour	pbs.org/newshour/
MICASE	Michigan corpus academic discourse	Informative	Classroom, seminar, and other campus interaction types.	hti.umich.edu/m/micase

(*continued*)

APPENDIX (*Continued*)

Abbreviation	Full Title	Register	Description	url (if available)
NPR	National Public Radio (also see: BUR) Boston Radio News	Informative	LDC corpus of rereadings of the news by Boston National Public Radio announcers	ldc.upenn.edu/Catalog/LDC96S36.html
NYT	*New York Times*	Informative	Web-downloaded material from the *New York Times* archives	nytimes.com;.ibiblio.org/slanews/internet/archives_other.html
NYRB	*New York Review of Books*	Informative	Web-downloaded from the *New York Review of Books*	nyrb.com
OTA	Oxford Text Archives	Varied	Print archives, including several genres	—
Prose	Literary prose and dialogue	Varied	Literature, generally where the entire book is downloadable	ota.ahds.ac.uk; hti.umich.edu; litrix.com
PD	Presidential Debates	Adversarial	Corpora from presidential archives	nara.gov; bushisms.com; sourcedocuments.com
Q/A	Presidential News Conferences	Informative	Corpora from presidential archives	as above
SCR	Segrin Couples' Remedial	Remedial?	Sound and transcript archives from the Segrin study	—
SCS	Segrin Couples' Supportive	Supportive	Sound and transcript files from the Segrin study's 'What do you like best about [your partner]?'	—
SWB	Switchboard	Supportive	LDC corpus of phone conversations	ldc.upenn.edu/Catalog/LDC97S62.html; ldc.upenn.edu/Catalog/LDC93S7.html
Tagliamonte	Tagliamonte & Smith	Interviews	Interviews of Britishers from different areas	—
T/L	Tripp/Lewinsky	Supportive?	Governmental/ABC corpus of phone conversations	abcnews.go.com/sections/us/dailynews/tripptapes981118.html
USC	Marketplace (also see: Mkt)	Informative	LDC corpus of NPR economic news program	ldc.upenn.edu